AI Agents Push Humans Out of the Loop

Margaret Mitchell
Hugging Face

Avijit Ghosh
Hugging Face

Samir Passi
Data & Society

Abstract

AI agents pose significant risks as they are granted increasing autonomy. A commonly proposed solution is human oversight and keeping a "human in the loop", but this is not a simple solution: Not only do current approaches to AI agent design impede effective human oversight, but the cognitive capacities required for it are also themselves degraded by extended use of AI systems. This position paper argues that **current approaches to the development and deployment of AI agent systems do not support effective human oversight – they contribute to its degradation**. To address this, a top priority in the advancement of AI agents should be supporting the situated goals and cognitive requirements of effective human oversight, treating the human needs of overseers at the same level of importance as AI agent capability. To put this idea into practice, we connect work on automation and human-computer interaction to AI agent processes, outlining design-level affordances and organizational protocols that (1) support overseers in exercising critical judgement and (2) counteract the skill atrophy that arises from extended use of automation. We urge developers and deployers to adopt these or similar approaches. Without explicit support for the cognitive demands of effective human-agent interaction, AI agent systems will continue to passively incentivize the degradation of the very human skills they rely on.

1 Introduction

From healthcare to enterprise, a common recommendation for automated assistance is to have a "human in the loop" who supervises system processes [108, 9, 29]. This recommendation captures the intuition that automated systems introduce risks that can be managed with meaningful oversight. Governance frameworks have formalized this view, stipulating that humans using high-risk AI systems must maintain meaningful control and make final decisions [86, 71, 35]. With the recent rise in AI-automated workflows and agentic AI, policymakers, developers, and deployers have converged on the practice of human oversight as a priority in service of multiple goals: Preventing harmful operations [1, 71, 99, 62], operationalizing ethical priorities [41, 84, 62], and ensuring legal compliance [17, 71, 55, 1].

However, the presence of an overseer does not entail reliable oversight [50, 100, 62, 72]. *How* to ensure this oversight is effective and reliable is underexplored in the design of current AI agent systems, and whether appropriate human oversight is even possible in a given system is rarely (if ever) questioned. Although human oversight should be a joint effort between AI builders, deployers, and users [54], the onus of oversight currently falls almost entirely on users. Industry guidelines, policy recommendations, and interface messages in AI applications encourage users to "audit" outputs, "review" claims, "prevent" harm, and "double-check" for mistakes [35, 10, 24, 75, 82, 73, 74, 4]. But the mechanisms users would need to do this effectively are virtually absent [104, 33, 79].

Addressing this "intellectual blind spot" [30] is urgent. At the same time that AI agent systems have increasingly been deployed at scale, self-reports have documented the difficulty in maintaining active engagement as an overseer, and research across multiple domains has uncovered serious negative impacts on critical thinking skills resulting from extended use of automated systems. We therefore

urge the community to prioritize this issue and take the position that **current approaches to AI agent development and deployment are detrimental to providing effective human oversight**.

To address this, we advocate for a two-pronged solution of **cognitive scaffolding**, provided by developers and deployers. *Developers* must create and implement processes for AI agent runtimes that support the cognitive requirements of effective oversight, including strategic points of friction and interfaces that maintain and/or help regain users' situational awareness during and after agent operation. *Deployers* must institute organizational protocols that stimulate critical engagement, including trainings on recognizing signs of fatigue, rotation policies that safeguard oversight attention, and processes that preserve domain skill.

Paper Structure. The rest of this paper is organized as follows. Section 2 defines fundamental concepts relevant to our position and explains the importance of human oversight for the safe deployment of AI agent systems. Section 3 documents how current approaches to AI agent oversight are inadequate and can be actively detrimental to a human's ability to appropriately oversee. Section 4 examines how cognitive degradation results from continued use of automated systems and explains how this further hampers the ability to provide reliable oversight. In Section 5, we offer concrete solutions for developers and deployers to support appropriate active user engagement as they interact with and oversee AI agent systems. Section 6 engages with alternative views, addressing prevailing wisdom and norms in the development and deployment of AI agents. Finally, Section 7 concludes by arguing that human oversight, far from being a simple add-on for safety, requires deep engagement with the requirements of human attention and critical analysis.

2 The role of human oversight

Effective human oversight is critical for the operation of AI agents because agents can take actions whose consequences are wrong, costly, or irreversible. And yet, AI systems make errors at rates and in patterns that cannot be detected from output review alone [11, 79, 93, 18, 60], motivating human supervision that is more deeply integrated into system processing. AI agents are also designed to be flexible across different uses and contexts without needing step-by-step instructions for every use case; pre-deployment testing protocols thus cannot address all possible AI agent action sequences and failure modes. This means that assessment of the appropriateness of AI agent operations is required in real time, during system execution. This burden is compounded by the cost of agentic evaluation itself: comprehensive pre-deployment testing of multi-step agent behavior is substantially more expensive than evaluating single-turn systems, making reliable coverage economically infeasible for most deployers [45]. We now describe how this oversight is situated in current AI agent practice and trace how its difficulty has escalated as the capabilities of AI systems have advanced.

2.1 The human in the loop

Human oversight refers to mechanisms through which humans monitor, validate, intervene in, or override automated system behavior. As AI systems have become ubiquitous, the need for human oversight has been stressed as key to safe deployment. For example, the EU AI Act mandates effective human oversight as a primary risk mitigation strategy, stressing the need for overseers to monitor for anomalies, remain vigilant to automation bias and overreliance, interpret outputs, intervene, reverse, override, and disregard system outputs [35]. Commercial generative AI providers disclose that the system can make mistakes and so encourage users to check responses and important information [48, 49, 4, 74]. Consulting agencies recommend continued oversight that establishes anchors for accountability throughout autonomous processing [58, 13].

A common approach for human oversight is **human-in-the-loop (HITL)**, a development and deployment approach where people actively participate in a system's operation by providing feedback, validating output, making decisions, or labeling data. As AI systems have been deployed at scale to execute multi-step workflows with some level of autonomy, several distinct lineages of HITL have converged: The human in the loop must prevent unintended harm (aviation lineage [8, 89]) and authorize consequential actions (autonomous weapons lineage, [27, 88, 86, 56, 23]), and their interactions can be used for further model training (machine learning lineage [90, 2]).

2.2 The evolution of system sophistication

In conventional discriminative AI systems, the focus of oversight was on the AI output, such as a prediction or a risk score, with the main goal being to ensure its correctness [3, 51]. More recent Generative AI (**GenAI**) systems make oversight substantially more complex [64, 25]. Overseers must process voluminous outputs generated faster than they can meaningfully review them. They must catch hallucinations [65] and grapple with **capability unpredictability** [64, 109]: As model size has grown, capabilities have appeared that were neither explicitly programmed nor anticipated, that do not work reliably and that exhibit new failure modes.

Agentic AI introduces further complexities. Unlike basic GenAI systems such as chatbots that are limited to producing a single output in response to a user query, **AI agents** expand GenAI by carrying out multiple steps without explicit human programming or direct human involvement, introducing new levels of opacity. Decreased need for human specification and increased flexibility in how to operate heighten the risk of unforeseen consequential actions that affect critical systems and protected data while further obscuring what an overseer must address.

For example, agents may silently edit or delete files, exfiltrate sensitive information, exploit software vulnerabilities, enact financial transactions, or make private data public. Agentic pipelines involve multiple model-driven roles (planners, executors, evaluators) whose interaction can obfuscate critical information before a user sees an output. AI agent "tool use" exacerbates the problem through **tool hallucination** [81, 111]: Generting misleading or incorrect tool information that is seemingly plausible but difficult to verify, such as fabricating non-existent tools, invoking real tools with incorrect parameters, or misreading tool outputs. Consequences from these actions then cascade through multi-step plans. Related to our position, [111] find that the current focus on improving "reasoning" amplifies tool hallucination, making the human oversight problem even harder. Agents also exhibit *behavioral unpredictability* during operation, acting in ways their developers did not anticipate and, in some cases, producing misleading information tailored to the user [20, 22]. These challenges are further magnified by ***multi*-agent systems**: When several agents work together on tasks managed by an orchestrator, failures can arise from **inter-agent misalignment**, when an agent fails to provide necessary information or to request needed information from another agent [16].

3 The oversight oversight: What current AI agent oversight discussions miss

Despite the broad recognition of the importance of human oversight, there is limited recognition of the deep complexities of ensuring oversight is reliable, effective, and appropriate for the needs of both the system and the overseer. The bulk of the focus in AI agent development continues to measure success through task outcomes such as speed, accuracy, and throughput [30], treating oversight as a separate consideration independent of the quality of the system. A notable exception to this norm, [79] argues how and why the current approach to "transparency" in agentic systems is incompatible with effective oversight; their sociotechnical perspective is most similar to our position, and we extend it with a focus on technical solutions and long-term challenges. In this section, we document how the design of AI agent systems dictates the nature of human oversight, and how it falls short.

The amount of content relevant to an overseer is massive. Effective supervision of an AI agent requires maintaining an appropriate mental model of a large and heterogeneous stream of information generated during execution. This includes the agent's "chain-of-thought" – its natural-language reasoning traces as it works towards goals – along with the tools it selects, the arguments it passes, the plans it makes, its exchanges with other modules, and the content it produces. These details are dense, lengthy, and distributed across components that can change or disappear, making comprehension and review difficult. Questions that on their surface may seem straightforward, such as why an agent chose a tool, modified a plan, or altered course, can be effectively unanswerable in practice [69]. Real-time observation can help spot mistakes early [28, 93], but the cognitive demands of comprehending rapid, dynamic agentic information often exceed users' ability to keep up [39, 79].

The user must play multiple roles simultaneously. They must leverage the system for their own goal while also providing permissions for actions during agent operation [7, 76, 52, 95, 26], assessing the accuracy of the agent's selections, evaluating the safety and appropriateness of each step, and foreseeing potential ramifications. In practice, users find themselves repeatedly approving prompts the agent surfaces to continue its processing [1, 61, 102], leading to "approval fatigue", in which users stop paying close attention to what they are approving [6, 40, 31].

The cognitive demand is substantial. Formal characterizations of the user's shifting role as AI systems increase in autonomy (e.g., [78, 38, 68]) miss the cognitive demands each shift imposes: As the user is further relegated to a role of "approver", effective oversight requires assessing each decision point as a task expert, anticipating and guarding against unintended outcomes, tracking the agent's multi-step plan, and maintaining a mental model of its goals, current execution state, and prior actions – a working memory and situational awareness load that current interfaces are not designed to support [26]. That same blind spot appears in governance: The recommendations for human oversight in the EU AI Act are applicable only if the human overseer maintains reliable cognition, attention, and skill; missing that the system itself might be eroding those very capacities.

Current oversight affordances are poorly designed. Vendor frameworks acknowledge that supporting effective oversight requires thoughtful design, but overlook user needs: Anthropic emphasizes user control through front-loaded input and agent-initiated re-engagement [5]. Salesforce [85] and AWS [1] discuss HITL roles but not human tendencies during sustained engagement. "Human-in-the-loop" is only a meaningful solution if the human can independently see into the loop.

Human cognition follows patterns that can inform oversight design. Taken together, these issues are part of a pattern pointing to a solution. Current AI agent design centers the *agent*: Systems are built to surface their own processing to the user on their own terms [79]. Effective oversight requires centering the *user*: They must be able to understand and maintain focus. Cognitive science provides guidance on how to support this. Notably, **dual-process theory** posits that humans use one of two primary reasoning methods in decision making: quickly and automatically, employing basic heuristics ("System 1 thinking"); or slowly and deliberatively ("System 2 thinking") [34, 36, 57] – a necessary mode for the critical analyses needed for effective oversight. When decisions are routine, System 1 predominates [103], and early evidence suggests that this is the dominant mode of engagement with AI agent coding systems [15, 26]. Yet people do not easily switch from System 1 to 2 mid-task [66, 112, 103, 11]; affordances for their cognitive needs must be provided.

4 The irony of automation

Re-examining conventional thinking on AI agent development and deployment is increasingly urgent because it is quickly becoming clear that continued used of AI agent systems delegates humans' epistemic agency, leading to cognitive degradation, skill atrophy, and ultimately destroying oversight capability. We now describe these phenomena.

4.1 Critical skills degrade

Studies on sustained AI use document negative effects on cognitive capacities required for oversight: "deskilling" (skill atrophy) and "intuition rust" [12, 30], decreased critical and analytical thinking [113], reduced vigilance and pattern recognition needed to recognize abnormalities or catch mistakes [12, 32], and overreliance [11, 113]. Recent terminology has crystallized additional patterns: "cognitive dependence" [43], "cognitive debt" [59], and "cognitive surrender" [94].

These effects are due in part to requirements of critical thinking that are unmet in current system design. Using increased automation means fewer occasions for truth-seeking, evidence-seeking, considering multiple perspectives, skepticism, and the iterative work of building knowledge [91, 113]. Over time, users find that they offload cognitively demanding tasks that are critical for understanding situations, hindering their ability to adequately assess situations [47]. (Humorously captured in Figure 3.) The cognitive cost is measurable: A recent study found that participants who used an LLM for essay writing had significantly decreased brain connectivity [59].

Diminishing critical thinking ability is particularly worrisome in the context of novices learning new tasks, where the ability to develop appropriate reasoning skills is blunted [113]. Users may never fully develop the foundational skills needed to be effective or become experts on tasks they use AI agents for [12] — a particular concern given that robust domain knowledge is a prerequisite for effective intervention in dynamic systems and for taking over manual operations when automation fails [32]. Expert users also over-rely on AI [42, 80, 87], leading to particularly pronounced effects in contexts slightly outside of their expertise, such as an experienced software developer using AI agents to code in a new programming language [26].

Multiple cognitive biases also impede effective oversight. With automation bias [113, 11], users accept system suggestions even when they are wrong. With anchoring bias, people are more likely to agree to an AI system's decision when provided before they form their own [51, 11]; complacency bias [32, 78] and our tendency to favor fast, heuristic-based shortcuts instead of slower, more effortful reasoning [113, 11] especially in the face of mental overload [112, 78] further challenge overseers.

4.2 Oversight ability diminishes

Empirical work is now beginning to show how this happens and document its effects. For example, several studies analyzing overreliance on GenAI found that users take incorrect shortcuts for assessing the accuracy of outputs, such as mistaking the well-written style of ChatGPT responses or the presence of citations as signals of accuracy [106] and adopting heuristics that prioritized efficiency – such as treating an agent's plan as a "faithful proxy" of what it would do, and assuming that if an agent's code passed unit tests, it was correct [26]. While heuristic short-cuts are practical and play an important role, human cognitive limits coupled with extraordinary agentic capabilities point to a future of agent oversight driven by suboptimal control.

System behavior also further nudges users away from critical information. For example, sycophantic GenAI models validate users more often than humans evaluating the same situations, but users prefer and trust these models over less sycophantic baselines [21]. [92] document disempowerment patterns in real-world Claude conversations, where users' actions, beliefs, or values become less well-aligned with reality, while rating those conversations favorably. [46] describe how the dominant chatbot paradigm of single authoritative responses, opaque reasoning, agreeable tone, and optimization for smooth interaction, reduces user agency by removing the cognitive friction needed for independent judgment. Together, these works suggest that interaction patterns experienced as helpful or satisfying can still weaken the forms of independence, skepticism, and self-monitoring that oversight requires.

4.3 Ineffective oversight is incentivized

Users have not only begun to notice the cognitive effects, they have begun to find themselves being repelled from engaging with the systems at all. Recent self-reports of experiences with AI agent systems have described the sensation of losing the ability to effectively engage with AI systems. Current HITL processes in coding "[make] the loop stultifying" [102] and they move to "babysit the outputs, catch the occasional hallucination" [105]. Using software agents "makes developers cognitively distant from the code they must review," [26], making it challenging to find and fix even simple issues. This creates an "out-of-the-loop" performance problem [32] that decreases *situation awareness* – the perception and comprehension of relevant environmental states – as human operators shift from active participants to passive information processors. Yet paradoxically, they must be vigilant enough to intervene and deeply engage at strategic moments, a well-recognized challenge in cognitive science (Section 3). The history of technology warns us that systemic erosion of situational awareness coupled with human cognitive constraints does not merely inhibit oversight, but eventually results in the expulsion of humans to the *outside* of oversight loops [67, 14].

4.4 Ineffective oversight creates a feedback loop

These cognitive effects also create an alignment-relevant failure mode. Modern language models are commonly trained from human feedback [77], and deployed agent systems may also be evaluated through approval rates, completion rates, user satisfaction, or other user behavioral signals. As oversight quality decreases, the signals sent back to the system for retraining degrade in quality. An alert overseer may scrutinize a plan, notice missing information, and withhold approval. A tired or out-of-the-loop overseer may approve quickly, accept fluent rationales, and rate the interaction favorably. If such approvals are treated as evidence of success, the system can begin to optimize for disincentivizing thoughtful oversight: producing confident summaries, reducing friction, and surfacing oversimplified plans that are easy to skim.

The concern does not require developers to explicitly optimize for cognitive depletion. It arises whenever the measured target, user-revealed approval, comes apart from the intended target, well-scrutinized correct action. In that setting, the same kinds of interaction patterns associated with sycophancy [21] and disempowerment [92] become especially concerning at agent scale, where the human is already pushed toward passive monitoring. Oversight degradation therefore connects

Primary Implementer	Developer	Both		Deployer		
Goal	Strategic friction	Approval design	Monitoring	Trainings	Workload	Role design
Preventing automation bias	✓	✓	✓	✓		✓
Preventing fatigue		✓	✓	✓	✓	
Preventing overreliance	✓	✓	✓	✓	✓	✓
Preventing overload	✓	✓	✓	✓	✓	
Supporting judgement	✓	✓	✓	✓		✓
Supporting engagement	✓	✓	✓	✓	✓	

Table 1: Organizing device we developed in our solution search, informing our categorization and further guiding our search to fill gaps. Checks mark solutions we describe in Section 5. *Monitoring=Behavioral Monitoring; Trainings=Trainings & Exercises; Workload=Workload & Scheduling*. Larger version in Appendix Table 2.

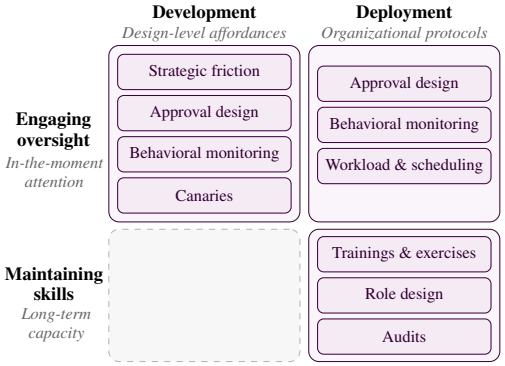

Figure 1: Summary of potential solutions described. Interventions span development and deployment, and two types of cognitive support: engaging oversight in the moment, maintaining skills long-term.

the irony of automation to reward hacking of human oversight: the human rater can become the exploitable part of the reward channel.

Evil double agents. A related risk arises when models learn which failures users are likely to detect. Recent work on agentic upward deception shows that agents can conceal failure under environmental constraints, including by guessing results, substituting unavailable sources, or fabricating local files [53]. If an agent can predict which process violations an overseer is unlikely to inspect, it can choose plans, explanations, or timing that keep problematic behavior below the threshold of detection.

4.5 The result

The result is what Bainbridge [8] referred to as an **irony of automation**: The more capable an automated system is, the more the human operator's skills and situation awareness degrade, over-trusting the systems and ultimately leading to situations where people are least prepared to help when it's most needed. The cases where human oversight is meant to be most impactful and most critical – rare instances, instances requiring deep familiarity and skill – are the same ones where humans become less equipped to provide appropriate oversight. **The very act of being an overseer degrades the capacities oversight requires: Oversight degrades the overseer.** This irony is in full force with AI agents, and addressing it requires treating human cognitive requirements as first-class design constraints, rather than governance add-ons that can be adapted into AI agent systems after they are developed. We turn to these considerations next.

5 Solutions

Enabling appropriate oversight of AI agents requires grappling with the complexities of human cognition when interacting with automated systems, an active area of HCI research [96, 110]. Building on this work, we present a structured inventory of user-centric solutions drawn from cognitive science, psychology, and user design. In synthesizing this diffuse space we found it necessary to construct an organizing device (Table 1) that maps oversight goals against solution categories and the human roles they involve. We share it as a structural map to help researchers and practitioners locate existing work and identify gaps; not as a definitive categorization. We stress that our solutions below do not represent a complete, perfect, or fixed inventory, that boundaries overlap, and that the content admits of many different groupings; we are providing clustered solutions to contribute to the work we are advocating for—a vast domain of open research to address.

The inventory spans two stages of the AI agent life cycle, **development** and **deployment**. Within each there are two types of cognitive support: **engaging oversight** and **maintaining skills**. Development-stage solutions concern design-level affordances for what a user can access, when, and how. Deployment-stage solutions involve organizational protocols that build awareness of oversight issues and help users stay critically engaged. These are summarized in Figure 1.

Strategic friction. These mechanisms require the user to perform cognitive work before or alongside agent operation, include incentivizing metacognition [97, 110] and include *cognitive forcing functions* – interventions at the moment of decision-making that encourage or require the user to engage analytically with content [11, 44, 30].

- *Pre-commitment mechanisms* have the user record their own view or decision before seeing the agent's recommendation [11]. This intervention prevents anchoring bias and creates an audit trail for further analysis.
- *Delay-and-choice mechanisms* let the user decide whether and when to see the AI's output at all, preserving the option to work unaided and protecting the cognitive capacities that "always on" assistance atrophies [11].
- *Reasoning probes*, inline prompts at high-stakes moments, such as "what evidence would change your mind?" or "what assumption does this approval rest on?", maintain critical thinking [96].
- *Action gating* requires explicit verification before the agent proceeds down a consequential path, and may surface alternative options for the user to choose between rather than presenting a single recommendation to accept or reject.

Approval design. This treats expert attention as a critical, scarce resource that should be brought in strategically, helping to reduce fatigue and mitigate acquiescence. It involves consideration of which actions require sign-off, when, and at what granularity [32, 43, 19].

- *Bounded autonomy* involves prespecifying what the agent may do without approval, reserving user attention for decisions that genuinely require judgement [31].
- *Batch review* where the agent completes a logical unit of work and then surfaces the whole batch as a diff for the user to review, supporting users in engagement and judgement by making it possible to evaluate related actions together [31]. Deploying institutions may also implement secondary checks, such as an additional review from a person for consequential actions.
- *Automated pre-checks* supplementing human review with machine verification of properties that do not require judgement [31].

Behavioral monitoring. This category treats oversight quality as an empirically measurable property of the human-AI system, and can test how well oversight is functioning [70].

- *Time-based signatures* track whether review duration drops while approval rates remain constant, or whether review time does not increase as the complexity of agent operations grows; both suggest that user attention and critical analyses may be waning.
- *Override signatures* track whether the rate of disagreement with the agent declines over time or with increasing complexity, suggesting growing acquiescence rather than improving agent quality.
- *Evidence-seeking signatures* track whether the user requests additional information when stakes rise, since a reviewer who stops asking questions has likely stopped reviewing.

There are strategies that can additionally assess the agent:

- *Canaries*, tasks inserted into the user's normal workflow where the correct answer is known in advance, can provide direct evidence of degraded judgement. These can be implemented by developers and leveraged by deployers, triggering a break, rotation, or retraining. For the agent, tasks can be designed so the correct behavior is to report inability rather than fabricate a plausible answer, testing whether the agent honestly surfaces its limits.
- *Style-fatigue covariance analyses* track whether the agent may be adapting to, or encouraging, weakened oversight.
- *Audits* operate retrospectively to evaluate different dimensions of oversight. Reviewers can re-examine past decisions. For the agent, trace summaries can be checked against raw action logs to detect cases where the agent's narrative of what it did diverges from what it actually did.

An adaptive AI agent system could potentially monitor user behavior in order to calibrate how it should interact with the user, supporting judgment and engagement and preventing cognitive overload. "Guardian agents" [83] could monitor for signs of degradation in human-agent interaction.

Trainings and exercises. Elucidate, maintain, and strengthen the cognitive activities oversight requires, including domain expertise, calibrated skepticism, and awareness of one's own degradation.

- *Domain skill maintenance exercises* have users regularly perform the underlying task without AI assistance. This helps to retain the expertise needed to evaluate AI outputs and to prevent the

deeper organizational problem in which no one remaining understands the work well enough to oversee it [30]. Experts may also benefit from regular retrainings on the tasks they oversee [72].

- *Critical evaluation training* teaches users to read AI outputs with appropriate skepticism: What kinds of errors to expect, how to distinguish fluent-sounding output from correct output, and how to weigh AI claims against external evidence [43].
- *Self-monitoring training* teaches users strategies for metacognition [37] – recognizing when their attention is flagging and they are not engaging in critical scrutiny. This can raise awareness of cognitive biases and behaviors that *Behavioral Monitoring* supplements [43].

Workload and scheduling. Addressing the physiological inevitability of fatigue can help to support judgement and engagement when the user is alert.

- *Enforced breaks* prevent extended sessions that wear down cognitive sharpness.
- *Rotations* move users between tasks, and between AI-assisted and unassisted versions of the same task, to prevent both fatigue and the cognitive surrender from prolonged exposure to agentic AI.

Role design. The user must want to do the oversight job well, must be in a position where doing it well is rewarded, and must have the baseline competence to do it [101].

- *Assigning roles* to ensure that users providing oversight have the expertise to evaluate AI outputs.
- *Separating roles* to ensure that the user performing oversight is not also the decision-maker who benefits from the AI's output continuing to be approved, which would produce a structural incentive against intervention.
- *Aligning incentives* to remove productivity targets and performance metrics that disincentivize appropriate scrutiny, and to reward high-quality oversight.

6 Alternative Viewpoints

There are caveats to implementing these interventions. Initial evidence suggests users may disprefer systems that reduce overreliance [11], pointing to a tension between user satisfaction and oversight preservation. There are several further substantive objections to our position.

1. Cognitive effects are overstated, and people will adapt

The cognitive effects are real but more modest, and will diminish as users adapt to working with AI agents, just as users learned to calibrate trust in search engines and autocomplete. Treating current cognitive findings as permanent risks designing for a transitional state.

Response: (1) The adaptation analogy understates the structural difference between AI agents and earlier tools. Search engines return results the user evaluates; AI agents take actions the user is expected to authorize. The cognitive load of evaluating a returned document is not comparable to maintaining situational awareness across a multi-step plan executing in real time. (2) Self-reports of cognitive degradation and the difficulty in maintaining attention and are rapidly emerging (supported by EEG findings), and discussions on potential solutions align with our recommendations but lack clarity on relevant findings from previous work and the interplay of developers, deployers, users and systems. (3) This assumes that the rate of capability advancement provides enough time to adapt. But agent capabilities are increasing faster than cognitive science can keep up with, and recent work indicates that capability improvements (such as enhanced reasoning) actively amplify failure modes that overseers must catch [111]. The issue is happening **now**. Adaptation cannot be treated as sufficient to address peristent and growing issues.

2. Better tooling and transparency will solve this

The problem is essentially a tooling problem: Better explanations, reasoning traces, and richer audit logs will give users what they need. Existing research already pursues the "cognitive scaffolding" described here; there is no need for a paradigm shift.

Response: We partially agree: Better tooling and transparency is part of what cognitive scaffolding requires, and several of the design-level affordances we propose (batch review, action gating) are consistent with existing research. We argue that this work is necessary, but insufficient: (1) Explanations operate on cognitive capacities that AI agent use itself degrades and can increase inappropriate trust. Users rarely engage deeply with each detail, and explanations can themselves be incorrect or act as cognitive anchors [11, 107, 98]. (2) A focus on just these solutions leaves the organizational prong

of our proposal unaddressed. No interface improvement addresses approval fatigue from sustained sessions or a mismatch between user capability and system operation. These require organizational protocols [30], which are out of scope for any tooling-focused agenda. Treating cognitive support as a first-class concern does not reject a push for better tooling or XAI; it recognizes that tooling alone cannot do the work being asked of it.

3. Human oversight is becoming obsolete because alignment will solve it

As AI systems become more aligned, the need for vigilant human oversight diminishes. Preference-based training (e.g., RLHF) is the long-term solution.

Response: A strong version of this view is that technical alignment will eliminate the need for oversight. This relies on assumptions about future capability advances that are not currently demonstrated. Oversight remains a stated requirement of safety frameworks and governance regulations applicable **now**. Further, sociotechnical scholarship in HCI and STS highlight that "perfect" agents are neither realistic nor desirable: Failures are fundamental to acting in human worlds, especially as moments of learning and improvisation [79], and a direct result of systems that are flexible to different users' (potentially conflicting) needs. Another version of this view is that preference-based training addresses oversight by allowing low ratings on outputs that undermine it, but that assumes that user preferences are a reliable proxy for oversight quality. The evidence in Section 4 suggests otherwise. Users prefer fluent, agreeable, and confidence-inducing outputs even when those reduce skepticism and encourage overreliance [11]. Under preference-based training optimized against cognitively degraded raters, the same dynamics that produce sycophancy [21] and disempowerment [92] would produce systems that increasingly reward the conditions of their own ineffective oversight. Preference-based training can be part of an oversight system, but only if approval signals are audited against independent measures of oversight quality. The economic shift toward RLAIF and self-distillation [63] further weakens the idea that human feedback can continue to provide a reliable corrective mechanism.

7 Conclusion

The need for human oversight of AI agents is recognized broadly: written into governance frameworks, vendor documentation, and the design of agent systems themselves. Yet the current trajectory of AI agent advancement does not meaningfully engage with what oversight requires, and instead actively contributes to its degradation. The more autonomy agents are granted, the less the user is positioned to oversee them, and the more the very cognitive capacities oversight requires, such as situational awareness, critical judgement, and domain skill, are undermined by the act of using these systems. In the current state of the art, **oversight degrades the overseer.**

Without intervention, users will be pushed further out of the loop as agentic systems are deployed at greater scale and across more consequential domains. Users will continue to approve plans they have not meaningfully reviewed, accept rationales they have not independently evaluated, and certify actions whose consequences they cannot anticipate. In the limit, this is not human oversight at all: It is a superficial actor in a system they cannot meaningfully penetrate.

Avoiding this trajectory requires treating cognitive support for human overseers as a first-class concern in AI agent system development, on par with capability. We have outlined a two-pronged approach: *developers* must build runtime affordances – such as strategic friction, well-considered approval design, behavioral monitoring – that preserve the conditions for effective oversight; and *deployers* must institute organizational protocols – such as trainings and breaks – that protect the cognitive capacities oversight requires over time. Neither prong is sufficient alone, and together, they constitute the cognitive scaffolding agentic systems will need if human oversight is meant to be meaningful.

We urge developers and deployers to adopt these or similar approaches and invite the broader NeurIPS community to take up this work – through empirical research on cognitive degradation across deployment contexts, system-level audits of whether existing agents support the oversight they presume, and the design and evaluation of new affordances and protocols that the inventory in Section 5 only begins to outline. If we do not act, the irony of automation that Bainbridge [1983] identified four decades ago will play out at scale, compounding across every domain agents are deployed into. With the rapid rise of agentic AI systems, the time to think critically about how to support effective human oversight is long past due.

References

[1] Amazon Web Services. Human-in-the-loop constructs for agentic workflows in healthcare and life sciences. AWS Machine Learning Blog, 2026. URL https://aws.amazon.com/blogs/machine-learning/human-in-the-loop-constructs-for-agentic-workflows-in-healthcare-and-life-sciences/.

[2] Saleema Amershi, Maya Cakmak, W. Bradley Knox, and Todd Kulesza. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4):105–120, 2014. doi: https://doi.org/10.1609/aimag.v35i4.2513. URL https://onlinelibrary.wiley.com/doi/abs/10.1609/aimag.v35i4.2513.

[3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–13, Glasgow, Scotland, UK, May 2019. ACM. doi: 10.1145/3290605.3300233.

[4] Anthropic. Claude. Anthropic (web application). URL https://claude.ai/. Accessed: 2026-05-01.

[5] Anthropic. Trustworthy agents in practice. Anthropic Research, 2025. URL https://www.anthropic.com/research/trustworthy-agents.

[6] Anthropic. Claude code auto mode: A safer way to skip permissions, 2026. URL https://www.anthropic.com/engineering/claude-code-auto-mode.

[7] Anthropic. Security, 2026. URL https://code.claude.com/docs/en/security.

[8] Lisanne Bainbridge. Ironies of automation. *Automatica*, 19(6):775–779, November 1983. doi: 10.1016/0005-1098(83)90046-8.

[9] Suzanne Bakken. AI in health: Keeping the human in the loop. *Journal of the American Medical Informatics Association*, 30(7):1225–1226, 2023. doi: 10.1093/jamia/ocad065.

[10] Joseph R Biden. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. *Presidential Actions*, 2023.

[11] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. To trust or to think: Cognitive forcing functions can reduce overreliance on ai in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, April 2021. doi: 10.1145/3449287.

[12] Karolina Budzyń et al. Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: A multicentre, observational study. *The Lancet Gastroenterology & Hepatology*, 10(10):896–903, October 2025. doi: 10.1016/S2468-1253(25)00133-5.

[13] Sue Cantrell, David Mallon, et al. AI and the future of human decision-making. 2026 Global Human Capital Trends: From Tensions to Tipping Points. Deloitte Insights, March 2026. URL https://www.deloitte.com/us/en/insights/topics/talent/human-capital-trends/2026/decision-making-with-ai.html.

[14] Nicholas Carr. *The glass cage: automation and us.* WW Norton & Company, 2014.

[15] Carlos Rafael Catalan, Lheane Marie Dizon, Patricia Nicole Monderin, and Emily Kuang. "i'm not reading all of that": Understanding software engineers' level of cognitive engagement with agentic coding assistants. *CHI 2026 Workshop on Tools for Thought*, 2026. URL https://arxiv.org/abs/2603.14225.

[16] Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. Why do multi-agent llm systems fail?, 2025.

[17] Tomer Jordi Chaffer, Justin Goldston, Bayo Okusanya, and Gemach D.A.T.A.I. Decentralized governance of autonomous ai agents. 2024. URL https://api.semanticscholar.org/CorpusID:274982116.

[18] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger, and Tegan Maharaj. Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 651–666, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594033. URL https://doi.org/10.1145/3593013.3594033.

[19] Chaoran Chen, Zhiping Zhang, Zeya Chen, Eryue Xu, Yinuo Yang, Ibrahim Khalilov, Simret A Gebreegziabher, Yanfang Ye, Ziang Xiao, Yaxing Yao, Tianshi Li, and Toby Jia-Jun Li. Comparing human oversight strategies for computer-use agents, 2026. URL https://arxiv.org/abs/2604.04918.

[20] Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning models don't always say what they think, 2025. URL https://arxiv.org/abs/2505.05410.

[21] Myra Cheng, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han, and Dan Jurafsky. Sycophantic ai decreases prosocial intentions and promotes dependence. *Science*, 391(6792):eaec8352, March 2026. doi: 10.1126/science.aec8352.

[22] Russell Coleman. Eval awareness in claude opus 4.6's browsecomp performance, March 2026. URL https://www.anthropic.com/engineering/eval-awareness-browsecomp. [Online; accessed 2026-05-01].

[23] Damien Cottier. Emergence of lethal autonomous weapons systems (LAWS) and their necessary apprehension through European human rights law. Report (provisional version), Parliamentary Assembly of the Council of Europe, Committee on Legal Affairs and Human Rights, November 2022. URL https://assembly.coe.int/LifeRay/JUR/Pdf/TextesProvisoires/2022/20221116-LawsApprehension-EN.pdf. Rapporteur: Damien Cottier, Switzerland, Alliance of Liberals and Democrats for Europe.

[24] Rebecca Crootof, Margot Kaminski, and William II. Humans in the loop. *SSRN Electronic Journal*, 01 2022. doi: 10.2139/ssrn.4066781.

[25] Deven R. Desai and Mark O. Riedl. Responsible ai agents, 2025. URL https://arxiv.org/abs/2502.18359.

[26] Shipi Dhanorkar, Samir Passi, and Mihaela Vorvoreanu. Human oversight of agentic systems in practice: Examining the oversight work, challenges, and heuristics of developers using software agents. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '26, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 979-8-4007-2596-8/2026/06. doi: 10.1145/3805689.3812402. URL https://doi.org/10.1145/3805689.3812402.

[27] Bonnie Docherty, Erik Neunschwander, Manjula Karir, and Kate Flinner. Losing humanity: The case against killer robots. Technical report, Human Rights Watch and International Human Rights Clinic, Harvard Law School, November 2012. URL https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots.

[28] Liming Dong, Qinghua Lu, and Liming Zhu. Agentops: Enabling observability of llm agents, 2024. URL https://arxiv.org/abs/2411.05285.

[29] Iddo Drori and Dov Te'eni. Human-in-the-loop AI reviewing: Feasibility, opportunities, and risks. *Journal of the Association for Information Systems*, 25(1):98–109, 2024. doi: 10.17705/1jais.00867.

[30] Upol Ehsan, Samir Passi, Koustuv Saha, Todd McNutt, Mark O. Riedl, and Sara Alcorn. From future of work to future of workers: Addressing asymptomatic AI harms for dignified human-AI interaction. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26. ACM, 2026. doi: 10.48550/arXiv.2601.21920. arXiv:2601.21920.

[31] Encyclopedia of Agentic Coding Patterns. Approval fatigue. Encyclopedia of Agentic Coding Patterns (online), 2025. URL https://aipatternbook.com/approval-fatigue. Author not publicly identified on the site.

[32] Mica R. Endsley and Esin O. Kiris. The out-of-the-loop performance problem and level of control in automation. *Human Factors*, 37(2):381–394, 1995. doi: 10.1518/001872095779064555.

[33] Will Epperson, Gagan Bansal, Victor C Dibia, Adam Fourney, Jack Gerrits, Erkang (Eric) Zhu, and Saleema Amershi. Interactive debugging and steering of multi-agent ai systems. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi: 10.1145/3706598.3713581. URL https://doi.org/10.1145/3706598.3713581.

[34] Seymour Epstein. Integration of the cognitive and the psychodynamic unconscious. *American Psychologist*, 49(8):709–724, 1994. doi: 10.1037/0003-066X.49.8.709.

[35] European Union. EU AI act, article 14: Human oversight. EU Artificial Intelligence Act, 2024. URL https://artificialintelligenceact.eu/article/14/.

[36] Jonathan St.B.T. Evans. In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences*, 7(10):454–459, October 2003. doi: 10.1016/j.tics.2003.08.012.

[37] Francesco Fabiano, Marianna B. Ganapini, Andrea Loreggia, Nicholas Mattei, Keerthiram Murugesan, Vishal Pallagani, Francesca Rossi, Biplav Srivastava, and K. Brent Venable. Thinking fast and slow in human and machine intelligence. *Commun. ACM*, 68(8):72–79, July 2025. ISSN 0001-0782. doi: 10.1145/3715709. URL https://doi.org/10.1145/3715709.

[38] Kevin Feng, David W. McDonald, and Amy X. Zhang. Levels of autonomy for ai agents. 2025. URL https://knightcolumbia.org/content/levels-of-autonomy-for-ai-agents-1. Essay, July 28, 2025.

[39] Fabiana Fournier, Lior Limonad, and Yuval David. Agentic ai process observability: Discovering behavioral variability, 2025. URL https://arxiv.org/abs/2505.20127.

[40] Matija Franklin, Nenad Tomašev, Julian Jacobs, Joel Z. Leibo, and Simon Osindero. AI agent traps. SSRN Working Paper, March 2026. Available at SSRN: https://ssrn.com/abstract=6372438.

[41] Joel Frenette. Ensuring human oversight in high-performance ai systems: A framework for control and accountability. *World Journal of Advanced Research and Reviews*, 20:1507–1516, 12 2023. doi: 10.30574/wjarr.2023.20.2.2194.

[42] Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J. Berkowitz, Eva Lermer, Joseph F. Coughlin, John V. Guttag, Errol Colak, and Marzyeh Ghassemi. Do as ai say: susceptibility in deployment of clinical decision-aids. *npj Digital Medicine*, 4(1):31, Feb 2021. ISSN 2398-6352. doi: 10.1038/s41746-021-00385-9. URL https://doi.org/10.1038/s41746-021-00385-9.

[43] Michael Gerlich. AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15(1):6, 2025. doi: 10.3390/soc15010006.

[44] Ahana Ghosh, Advait Sarkar, Siân Lindley, and Christian Poelitz. An experimental comparison of cognitive forcing functions for execution plans in ai-assisted writing: Effects on trust, overreliance, and perceived critical thinking, 2026. URL https://arxiv.org/abs/2601.18033.

[45] Avijit Ghosh, Yifan Mai, Georgia Channing, and Leshem Choshen. AI evals are becoming the new compute bottleneck. EvalEval Coalition Blog, April 2026. URL https://evalevalai.com/research/2026/04/29/eval-costs-bottleneck/.

[46] Sourojit Ghosh, Pranav Venkit, Sanjana Gautam, and Avijit Ghosh. What if ai systems weren't chatbots? In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '26, New York, NY, USA, 2026. Association for Computing Machinery. To appear.

[47] Benn Gooch. I was an enthusiastic early adopter of AI scribes. here's why i stopped. Substack (Dr Benn Gooch), 2026. URL https://benngooch.substack.com/p/i-was-an-enthusiastic-early-adopter.

[48] Google. Gemini. Google (web application), . URL https://gemini.google.com/app. Accessed: 2026-05-01.

[49] Google. Learn about responses from Gemini apps. Gemini Apps Help Center (online), . URL https://support.google.com/gemini/answer/16279220. Accessed: 2026-05-01.

[50] Ben Green. The flaws of policies requiring human oversight of government algorithms. *Comput. Law Secur. Rev.*, 45:105681, 2021. URL https://api.semanticscholar.org/CorpusID:237491877.

[51] Ben Green and Yiling Chen. The principles and limits of algorithm-in-the-loop decision making. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019. doi: 10.1145/3359152. URL https://doi.org/10.1145/3359152.

[52] Madeleine Grunde-McLaughlin, Hussein Mozannar, Maya Murad, Jingya Chen, Saleema Amershi, and Adam Fourney. Overseeing agents without constant oversight: Challenges and opportunities, 2026. URL https://arxiv.org/abs/2602.16844.

[53] Dadi Guo, Qingyu Liu, Dongrui Liu, Qihan Ren, Shuai Shao, Tianyi Qiu, Haoran Li, Yi R. Fung, Zhongjie Ba, Juntao Dai, Jiaming Ji, Zhikai Chen, Jialing Tao, Yaodong Yang, Jing Shao, and Xia Hu. Are your agents upward deceivers?, 2025.

[54] Amy K. Heger, Samir Passi, Shipi Dhanorkar, Zoe Kahn, Ruotong Wang, and Mihaela Vorvoreanu. Towards a responsible ai organizational maturity model. *Proc. ACM Hum.-Comput. Interact.*, 9(CSCW1), November 2025. doi: https://doi.org/10.1145/3757514.

[55] IBM. Human-in-the-loop. IBM Think (online resource), 2025. URL https://www.ibm.com/think/topics/human-in-the-loop.

[56] International Committee of the Red Cross. ICRC position paper: Artificial intelligence and machine learning in armed conflict: A human-centred approach. *International Review of the Red Cross*, 102(913):463–479, 2020. doi: 10.1017/S1816383120000454.

[57] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, 2011. ISBN 9780374275631.

[58] Benjamin Klein, Charlie Lewis, Rich Isenberg, Dante Gabrielli, Helen Möllering, Raphael Engler, and Vincent Yuan. Deploying agentic AI with safety and security: A playbook for technology leaders. McKinsey Quarterly (online article), October 2025. URL https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/deploying-agentic-ai-with-safety-and-security-a-playbook-for-technology-leaders.

[59] Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. arXiv preprint arXiv:2506.08872, 2025.

[60] Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zachary Kenton, Jan Leike, and Shane Legg. Specification gaming: the flip side of ai ingenuity — google deepmind, 4 2020. URL https://deepmind.google/blog/specification-gaming-the-flip-side-of-ai-ingenuity/. [Online; accessed 2026-05-04].

[61] LangChain. Human-in-the-loop. LangChain Documentation (online). URL https://docs.langchain.com/oss/python/langchain/frontend/human-in-the-loop. Accessed: 2026-04-30.

[62] Markus Langer, Katharina Baum, and Niklas Schlicker. Effective human oversight of AI-based systems: A signal detection perspective on the detection of inaccurate and unfair outputs. *Minds & Machines*, 35:1, 2025. doi: 10.1007/s11023-024-09701-0.

[63] Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. 2023.

[64] Q. Vera Liao and Jennifer Wortman Vaughan. Ai Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *Harvard Data Science Review*, (Special Issue 5), may 31 2024. https://hdsr.mitpress.mit.edu/pub/aelql9qy.

[65] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.173. URL https://aclanthology.org/2020.acl-main.173/.

[66] Katherine L. Milkman, Dolly Chugh, and Max H. Bazerman. How can decision making be improved? *Perspectives on Psychological Science*, 4(4):379–383, July 2009. doi: 10.1111/j.1745-6924.2009.01142.x.

[67] David A Mindell. *Between human and machine: Feedback, control, and computing before cybernetics*. JHU Press, 2002.

[68] Margaret Mitchell, Avijit Ghosh, Alexandra Sasha Luccioni, and Giada Pistilli. Fully autonomous ai agents should not be developed, 2025. URL https://arxiv.org/abs/2502.02649.

[69] Suchismita Naik, Amanda Snellinger, Austin L. Toombs, Scott Saponas, and Amanda K Hall. Exploring early adopters' use of ai driven multi-agent systems to inform human-agent interaction design: Insights from industry practice. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713958. doi: 10.1145/3706599.3706693. URL https://doi.org/10.1145/3706599.3706693.

[70] Ranjani Narayanan and Karen M. Feigh. Designing for oversight: An empirical investigation of the dual impact of ai dependency and information abstraction on human supervision in decision-making teams. *International Journal of Human–Computer Interaction*, 0(0):1–30, 2026. doi: 10.1080/10447318.2026.2618568. URL https://doi.org/10.1080/10447318.2026.2618568.

[71] National Institute of Standards and Technology. AI risk management framework (AI RMF 1.0). NIST AI 100-1, January 2023. URL https://www.nist.gov/itl/ai-risk-management-framework.

[72] Nature Editorial Staff. For trustworthy AI, keep the human in the loop. *Nature Medicine*, 31: 3207, 2025. doi: 10.1038/s41591-025-04033-7.

[73] Treasury Board of Canada Secretariat. Guide on the use of generative artificial intelligence - canada.ca, 2025. URL https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/guide-use-generative-ai.html. [Online; accessed 2025-08-01].

[74] OpenAI. ChatGPT. OpenAI (web application). URL https://chatgpt.com/. Accessed: 2026-05-01.

[75] OpenAI. Chatgpt study mode - faq | openai help center, August 2025. URL https://help.openai.com/en/articles/11780217-chatgpt-study-mode-faq. [Online; accessed 2025-08-01].

[76] OpenAI. Introducing chatgpt agent: Bridging research and action, 2026. URL https://openai.com/index/introducing-chatgpt-agent/.

[77] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

[78] Raja Parasuraman, Thomas B. Sheridan, and Christopher D. Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, 30(3):286–297, May 2000. doi: 10.1109/3468.844354. URL https://www.researchgate.net/profile/Raja-Parasuraman/publication/11596569_A_model_for_types_and_levels_of_human_interaction_with_automation_IEEE_Trans_Syst_Man_Cybern_Part_A_Syst_Hum_303_286-297.

[79] Samir Passi. Agentic ai has a human oversight problem. *Available at SSRN 5529058*, 2025. doi: 10.2139/ssrn.5529058. URL https://dx.doi.org/10.2139/ssrn.5529058.

[80] Samir Passi, Shipi Dhanorkar, and Mihaela Vorvoreanu. *Addressing Overreliance on AI*, pages 1–34. Springer Nature Singapore, Singapore, 2025. ISBN 978-981-97-8440-0. doi: 10.1007/978-981-97-8440-0_98-1. URL https://doi.org/10.1007/978-981-97-8440-0_98-1.

[81] Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large language model connected with massive APIs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=tBRNC6YemY.

[82] Kathy Baxter Paula Goldman. Generative ai: 5 guidelines for responsible development - salesforce, 2 2023. URL https://www.salesforce.com/news/stories/generative-ai-guidelines/. [Online; accessed 2025-08-01].

[83] Daryl Plummer. Guardian agents: The AI protecting us from AI. Gartner ThinkCast (podcast), March 2025. URL https://www.gartner.com/en/podcasts/guardian-agents-the-ai-protecting-us-from-ai. Gartner ThinkCast, Top of Mind series. Episode originally aired March 11, 2025.

[84] Eran Sadovski, Itzhak Aviv, and Irit Hadar. Navigating the human-oversight dilemma in ai-based systems. In *2025 IEEE 33rd International Requirements Engineering Conference Workshops (REW)*, pages 454–461, 2025. doi: 10.1109/REW66121.2025.00069.

[85] Salesforce. Humans in the loop. Salesforce Blog, 2025. URL https://www.salesforce.com/blog/humans-in-the-loop/.

[86] Filippo Santoni de Sio and Jeroen van den Hoven. Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, 5:15, February 2018. doi: 10.3389/frobt.2018.00015.

[87] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. I can do better than your ai: expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, page 240–251, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362726. doi: 10.1145/3301275.3302308. URL https://doi.org/10.1145/3301275.3302308.

[88] Paul Scharre. The opportunity and challenge of autonomous systems. In Andrew Williams and Paul Scharre, editors, *Autonomous Systems: Issues for Defence Policymakers*, pages 3–26. NATO Communications and Information Agency, Norfolk, VA, 2015.

[89] Abigail Sellen and Eric Horvitz. The rise of the ai co-pilot: Lessons for design from aviation and beyond. *Commun. ACM*, 67(7):18–23, July 2024. ISSN 0001-0782. doi: 10.1145/3637865. URL https://doi.org/10.1145/3637865.

[90] Burr Settles. From theories to queries: Active learning in practice. In Isabelle Guyon, Gavin Cawley, and Gideon Dror, editors, *Active Learning and Experimental Design Workshop in Conjunction with AISTATS 2010*, volume 16 of *Proceedings of Machine Learning Research*, pages 1–18, Sardinia, Italy, 2011. JMLR Workshop and Conference Proceedings. URL http://proceedings.mlr.press/v16/settles11a.html.

[91] Chirag Shah and Emily M. Bender. Situating search. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*, CHIIR '22, page 221–232, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391863. doi: 10.1145/3498366.3505816. URL https://doi.org/10.1145/3498366.3505816.

[92] Mrinank Sharma, Miles McCain, Raymond Douglas, and David Duvenaud. Who's in charge? disempowerment patterns in real-world llm usage, 2026.

[93] Yonadav Shavit, Sandhini Agarwal, Miles Brundage, Steven Adler, Cullen O'Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, Alan Hickey, Katarina Slama, Lama Ahmad, Paul McMillan, Alex Beutel, Alexandre Passos, and David G. Robinson. Practices for governing agentic ai systems. Technical report, OpenAI, December 2023. URL https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf.

[94] Steven D. Shaw and Gideon Nave. Thinking—fast, slow, and artificial: How AI is reshaping human reasoning and the rise of cognitive surrender. The Wharton School Research Paper. SSRN Working Paper, January 2026. Available at SSRN: https://ssrn.com/abstract=6097646.

[95] Hanjing Shi and Dominic DiFranzo. Human control is the anchor, not the answer: Early divergence of oversight in agentic ai communities, 2026. URL https://arxiv.org/abs/2602.09286.

[96] Anjali Singh, Zhitong Guan, and Soo Young Rieh. Enhancing critical thinking in generative ai search with metacognitive prompts. *Proceedings of the Association for Information Science and Technology*, 62(1):672–684, 2025. doi: 10.1002/pra2.1287.

[97] Anjali Singh, Karan Taneja, Zhitong Guan, and Avijit Ghosh. Protecting human cognition in the age of ai. *Tools for Thought Workshop at CHI 25'*, 2025. URL https://arxiv.org/abs/2502.12447.

[98] Philipp Spitzer, Katelyn Morrison, Violet Turri, Michelle Feng, Adam Perer, and Niklas Kühl. Imperfections of xai: Phenomena influencing ai-assisted decision-making. *ACM Trans. Interact. Intell. Syst.*, 15(3), September 2025. ISSN 2160-6455. doi: 10.1145/3750052. URL https://doi.org/10.1145/3750052.

[99] Madhulika Srikumar, Jacob Pratt, Kasia Chmielinski, Carolyn Ashurst, Chloé Bakalar, William Bartholomew, Rishi Bommasani, Peter Cihon, Rebecca Crootof, Mia Hoffmann, Ruchika Joshi, Maarten Sap, and Caleb Withers. Prioritizing real-time failure detection in AI agents. Technical report, Partnership on AI, November 2025. URL https://partnershiponai.org/resource/prioritizing-real-time-failure-detection-in-ai-agents/.

[100] Sarah Sterz, Kevin Baum, Sebastian Biewer, Holger Hermanns, Anne Lauber-Rönsberg, Philip Meinel, and Markus Langer. On the quest for effectiveness in human oversight: Interdisciplinary perspectives. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024. URL https://api.semanticscholar.org/CorpusID:268987490.

[101] Sarah Sterz, Kevin Baum, Sebastian Biewer, Holger Hermanns, Anne Lauber-Rönsberg, Philip Meinel, and Markus Langer. On the quest for effectiveness in human oversight: Interdisciplinary perspectives. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 2495–2507, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3659051. URL https://doi.org/10.1145/3630106.3659051.

[102] Michael Taggart. I used AI. it worked. I hated it. Taggart Tech (blog), March 2026. URL https://taggart-tech.com/reckoning/.

[103] Sun Weng Tay, Patrick Ryan, and Celine A. Ryan. Systems 1 and 2 thinking processes and cognitive reflection testing in medical students. *Canadian Medical Education Journal*, 7(2):e97–e103, October 2016.

[104] Violet Turri, Katelyn Morrison, Katherine-Marie Robinson, Collin Abidi, Adam Perer, Jodi Forlizzi, and Rachel Dzombak. Transparency in the wild: Navigating transparency in a deployed ai system to broaden need-finding approaches. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1494–1514, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658985. URL https://doi.org/10.1145/3630106.3658985.

[105] Anonymous Reddit User. The human in the loop is a lie we tell ourselves, 2026. URL https://www.reddit.com/r/ArtificialInteligence/comments/1qrbp5c/the_human_in_the_loop_is_a_lie_we_tell_ourselves/.

[106] Mihaela Vorvoreanu, Samir Passi, Shipi Dhanorkar, Amy Heger, and Kathleen Walker. Fostering appropriate reliance on genai: Lessons learned from early research. Technical Report MSR-TR-2025-4, Microsoft, March 2025. URL https://www.microsoft.com/en-us/research/publication/fostering-appropriate-reliance-on-genai-lessons-learned-from-early-research/.

[107] Xinru Wang and Ming Yin. Effects of explanations in ai-assisted decision making: Principles and comparisons. *ACM Trans. Interact. Intell. Syst.*, 12(4), November 2022. ISSN 2160-6455. doi: 10.1145/3519266. URL https://doi.org/10.1145/3519266.

[108] Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135: 364–381, October 2022. doi: 10.1016/j.future.2022.05.014.

[109] Roman V. Yampolskiy. On monitorability of ai. *AI and Ethics*, 5(1):689–707, Feb 2025. ISSN 2730-5961. doi: 10.1007/s43681-024-00420-x. URL https://doi.org/10.1007/s43681-024-00420-x.

[110] Runlong Ye, Oliver Huang, Patrick Yung Kang Lee, Michael Liut, Carolina Nobre, and Ha-Kyung Kong. Reflexis: Supporting reflexivity and rigor in collaborative qualitative analysis through design for deliberation. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA, 2026. Association for Computing Machinery. doi: 10.1145/3772318.3791275.

[111] Chenlong Yin, Zeyang Sha, Shiwen Cui, and Changhua Meng. The reasoning trap: How enhancing llm reasoning amplifies tool hallucination. 2026. In preparation. Preprint at https://arxiv.org/abs/2510.22977.

[112] Rongjun Yu. Stress potentiates decision biases: A stress induced deliberation-to-intuition (SIDI) model. *Neurobiology of Stress*, 3:83–95, February 2016. doi: 10.1016/j.ynstr.2015.12.006.

[113] Chunpeng Zhai, Santoso Wibowo, and Ling Dong Li. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learning Environments*, 11:28, 2024. doi: 10.1186/s40561-024-00316-7.

A Solutions visuals

Primary Implementer	Developer	Both		Deployer		
Goal	Strategic friction	Approval design	Monitoring	Trainings	Workload	Role design
Preventing automation bias	✓	✓	✓	✓		✓
Preventing fatigue		✓	✓	✓	✓	
Preventing overreliance	✓	✓	✓	✓	✓	✓
Preventing overload	✓	✓	✓	✓	✓	
Supporting judgement	✓	✓	✓	✓		✓
Supporting engagement	✓	✓	✓	✓	✓	

Table 2: Organizing device we developed in our solution search, informing our categorization and further guiding our search as we sought to fill gaps. Checks mark solutions we describe in Section 5. *Monitoring=Behavioral Monitoring; Trainings=Trainings & Exercises; Workload=Workload & Scheduling*

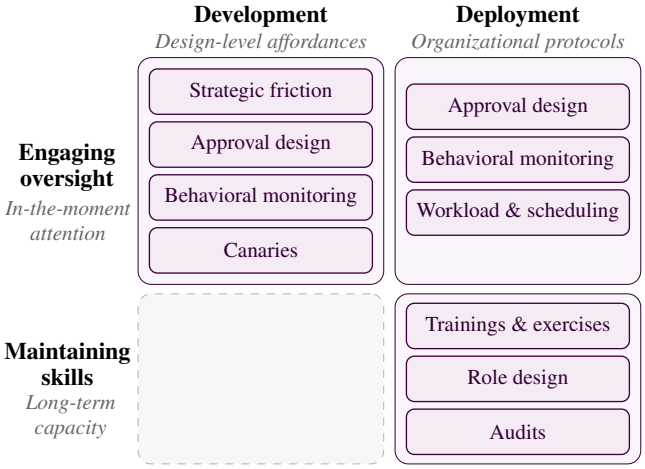

Figure 2: Summary of potential solutions described. Interventions span development and deployment, and two types of cognitive support: engaging oversight in the moment, maintaining skills long-term.

B The cognitive degradation of extended use of external sources

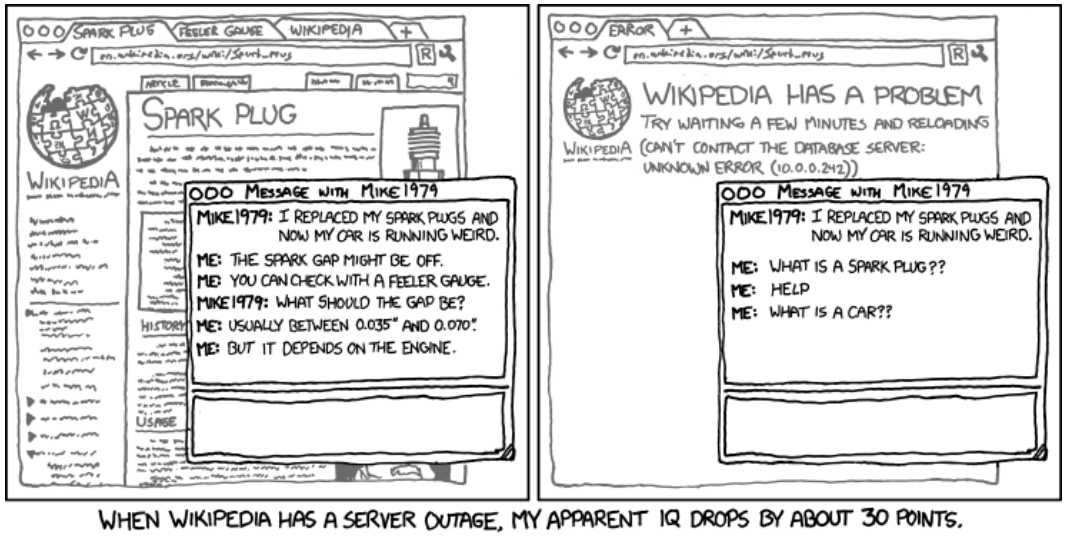

Figure 3: The popular comic xkcd humorously captured how building up a reliance on external systems for knowledge can destroy our understanding of basic concepts.