Attention-Constrained Oversight for Agentic AI

Risk, Reversibility, Authority, and Human Attention

Johnny Kao

Independent Researcher, Tokyo, Japan

ORCID: https://orcid.org/0009-0004-1446-6001

Working Paper | Version 1.2 | 24 September 2026

Abstract

As artificial intelligence systems move from recommending actions to executing them, human oversight can no longer be treated as a binary choice between autonomy and a human-in-the-loop. This paper proposes an attention-constrained framework for designing agent autonomy around four variables: risk, reversibility, authority, and human attention. Risk, reversibility, and authority describe the consequence profile of an action; human attention describes the scarce oversight resource available to review it. Drawing on research on levels of automation, appropriate reliance, automation bias, meaningful human oversight, and recent empirical studies of software agents, the paper argues that frequent approval requests can reduce rather than increase meaningful control when they consume attention faster than humans can exercise judgment. The framework introduces an action-level oversight demand, an explicit human-attention budget, a lightweight extension for cumulative exposure across action sequences, and two design principles: exception-based oversight and approval quality over approval count. It further distinguishes four complementary modes of oversight - boundary setting, monitoring, escalation, and audit - and shows how they can be combined so that routine, bounded actions proceed without synchronous review while consequential or difficult-to-reverse actions receive concentrated human attention. The paper also links these modes to task-level delegation contracts and proposes adaptive safeguards when review conditions appear to degrade. The contribution is integrative: reliable human oversight is framed not as maximizing human presence, but as allocating human judgment where its marginal value is highest.

Keywords: agentic AI; attention-constrained oversight; human oversight; human-in-the-loop; autonomy; approval fatigue; automation bias; human attention

Author note: This working paper develops and formalizes the autonomy-and-oversight framework first introduced in Kao (2026).

1. Introduction

The phrase human-in-the-loop has become a default answer to the governance problem created by increasingly autonomous AI systems. When an AI agent can write code, call tools, change external state, communicate with users, or continue through a multi-step workflow, placing a human approval point before consequential actions appears to preserve control. The intuition is reasonable: if the machine can act, a person should remain able to stop it.

The difficulty is that human presence is not the same as human control. An approval step has value only if the reviewer understands what is being proposed, has enough context to judge its consequences, and is willing and able to stop the action. When approval requests become frequent, repetitive, or low-value, the review process can degrade into a routine interaction rather than a meaningful decision. Research on

automation has long documented misuse, disuse, complacency, and the importance of appropriate reliance (Parasuraman & Riley, 1997; Lee & See, 2004). More recent human-AI research shows that explanations alone often do not produce calibrated reliance and may fail when users cannot independently verify the system's recommendation (Schemmer et al., 2023; Fok & Weld, 2024; Romeo & Conti, 2026).

Agentic AI makes this old problem operationally more severe. A decision-support system may present one recommendation at a time. An agent can generate dozens or hundreds of action opportunities inside a single task. Anthropic reported in 2026 that Claude Code users approved roughly 93 percent of permission prompts and explicitly linked high prompt volume to approval fatigue. Its engineering response combined automated risk classification with containment rather than relying on more human clicking (Anthropic, 2026). This is a useful empirical signal: adding approvals can consume the very attention on which those approvals depend.

Recent scholarship is also moving beyond a single approval-loop model. Zhu et al. (2026) argue that meaningful oversight should preserve human evaluative agency rather than reduce humans to rubber stamps. Wittens et al. (2026) situate oversight across task, monitoring, decision, organizational, and contextual layers. Dhanorkar et al. (2026), based on interviews with experienced developers using software agents, identify multiple forms of oversight in practice, including a priori control, co-planning, real-time monitoring, and post hoc review. These works establish that oversight is distributed across time and system layers rather than located at one approval gate.

This paper develops a complementary action-level question: given finite human attention, which agent actions should receive synchronous human judgment, which should be governed by policy, which should be monitored, and which can be audited after execution? The proposed framework centers four variables - risk, reversibility, authority, and human attention. Its contribution is not to claim that any variable is individually new. It is to connect consequence-based autonomy design with the scarcity of human attention and to treat oversight as an allocation problem rather than a binary architectural label.

2. Human-in-the-Loop Is a Routing Choice, Not a Safety Property

Classic research on automation already rejected the idea that systems should be classified simply as manual or automatic. Parasuraman, Sheridan, and Wickens (2000) described multiple types and levels of automation, emphasizing that the appropriate level depends on system performance, human workload, and the consequences of errors. The same principle applies to agentic AI, but the unit of analysis shifts from a stable system function to a stream of heterogeneous actions inside a workflow.

A coding agent, for example, may read files, edit a local branch, install a package, open a pull request, alter a production configuration, and deploy a service during the same session. Treating every action as equally review-worthy ignores the differences between them. Treating the entire session as autonomous ignores them in the opposite direction. The design problem is therefore not whether a human is 'in the loop' in the abstract, but how actions are routed among autonomous execution, policy controls, human escalation, and later audit.

This distinction also clarifies why oversight quality and oversight frequency are not interchangeable. A workflow can contain many human checkpoints while providing little meaningful control if the reviewer lacks time, context, or a credible veto. Conversely, a system with few synchronous approvals may preserve stronger human control if it constrains the agent's authority ex ante, monitors meaningful signals, escalates exceptions, and audits outcomes. Human-in-the-loop is therefore better treated as one possible routing mechanism within a broader oversight architecture, not as evidence that the architecture is safe by itself.

3. Related Literature and the Allocation Gap

3.1 Appropriate reliance and automation bias

The human-factors literature provides a mature foundation for understanding failures of oversight. Parasuraman and Riley (1997) distinguish appropriate use of automation from misuse and disuse, while Lee

and See (2004) frame trust as a mechanism that should support appropriate reliance rather than maximal reliance. In AI-assisted decision making, Schemmer et al. (2023) formalize appropriate reliance as the ability to accept useful AI advice while rejecting incorrect advice. Buçinca et al. (2021) show that cognitive forcing can reduce overreliance, but at a usability cost, illustrating that stronger human engagement is not free.

Recent syntheses reinforce the same tension. Passi, Dhanorkar, and Vorvoreanu (2024) review research on generative AI and conclude that both overreliance and underreliance can impair human-AI performance. Romeo and Conti (2026), reviewing 35 peer-reviewed studies, identify automation bias as a function of multiple factors including trust, expertise, task demands, cognitive workload, and the design of explanations. Fok and Weld (2024) argue that explanations improve decisions primarily when they help users verify whether an AI recommendation is correct; explanations that merely expose or narrate reasoning often do not provide that capability.

These findings matter for agent oversight because a permission prompt is itself a reliance decision. The user must decide whether to accept the agent's proposed action, often under time pressure and without independently reconstructing the full reasoning chain. If prompt volume rises while verifiability remains low, the formal presence of a human reviewer can coexist with weak substantive review.

3.2 Meaningful oversight in agentic systems

Contemporary oversight research increasingly recognizes that meaningful control must be designed rather than asserted. Zhu et al. (2026) distinguish AI operative agency from human evaluative agency and warn against oversight patterns that either eliminate useful autonomy or reduce humans to rubber stamps. Wittens et al. (2026) propose a multi-layer Human Oversight Anchor Framework spanning task, monitoring, decision, organizational, and broader contextual levels. Dhanorkar et al. (2026) add empirical evidence from software-agent users, finding that oversight work occurs before, during, and after execution rather than only at moments of intervention.

The remaining allocation gap is narrower but operationally important. These perspectives establish that oversight should be context-sensitive and temporally distributed, yet practitioners still need a compact rule for deciding which individual actions deserve scarce synchronous human attention. The framework developed below addresses that decision by linking action consequences to an explicit attention budget.

4. The Attention-Constrained Oversight Framework

Definition 1 (Action consequence profile). For an agent action i, the action consequence profile is described by risk R_i, reversibility V_i, and authority A_i. Risk asks how costly a wrong action could be; reversibility asks how easily the effects can be undone; authority asks what external state, resources, or rights the agent is able to change.

The three variables are deliberately distinct. A low-probability event can still justify strong controls when the downside is extreme. A moderately risky action may require less synchronous review if it is fully reversible and tightly contained. A highly capable model does not necessarily create a high-authority action if the surrounding system limits what the model can touch. Authority is therefore a property of the deployment and task, not merely of the model.

Table 1. Variables in the attention-constrained oversight framework

Variable	Operational question	Effect on oversight demand
Risk	How bad can a wrong action be?	Higher potential consequence increases oversight demand.

Variable	Operational question	Effect on oversight demand
Reversibility	How easily can the action be undone or restored?	Greater reversibility generally reduces the need for synchronous intervention.
Authority	What external state, resources, permissions, or commitments can the agent change?	Broader or more consequential authority increases oversight demand.
Human attention	How much high-quality review capacity is actually available?	Constrains how many actions can receive meaningful human review.

Let O_i denote the qualitative oversight demand for action i. The framework does not claim a universal numeric scoring function; instead it specifies directional relationships: oversight demand should increase with risk and authority and decrease as reversibility improves.

$$O_i = f(R_i, A_i, V_i)$$

In this expression, the function f is context-specific rather than empirically calibrated here. Its purpose is to make the design logic explicit: identical model confidence does not imply identical autonomy when the consequences, reversibility, or delegated authority differ.

Definition 2 (Human attention budget). Let B_H denote the finite amount of human review capacity available during a workflow or operating period. If h_i is the attention required to meaningfully review action i, then a feasible oversight design must respect the attention constraint below.

The attention cost of a review also depends on the review task itself. Context complexity and verifiability are two important drivers: complex context increases the material a reviewer must reconstruct, while high verifiability reduces the effort required to check a proposed action against independent evidence or clear acceptance criteria. The relationship can be represented conceptually as follows.

$$h_i = g(K_i, Q_i)$$

$$\Sigma h_i \leq B_H$$

This constraint is conceptually simple but changes the objective. If every routine action consumes review effort, fewer resources remain for unusual, ambiguous, irreversible, or high-authority decisions. Human attention is therefore not an unlimited safety layer. It is a scarce governance resource that must be allocated.

Definition 3 (Attention-constrained oversight). Human oversight is attention-constrained when review capacity is allocated preferentially to actions whose consequence profiles exceed an explicit autonomy boundary, rather than being consumed uniformly by routine actions.

Definition 4 (Cumulative exposure). Action-level routing should not assume that actions are independent. Let each action contribute an incremental exposure term. A sequence of individually low-consequence or technically reversible actions can accumulate into a trajectory whose aggregate consequence is no longer low.

$$C_t = \sum_{i=1}^{t} e_i$$

A policy can therefore permit individual actions while escalating when cumulative exposure crosses a domain-specific threshold. This is not a calibrated risk model; it is a lightweight safeguard against treating

repeated actions, information release, configuration changes, or resource commitments as independent events.

5. From Approval by Event to Control by Policy

A common implementation of human oversight is approval by event: each time an agent attempts an action outside a default permission set, a prompt is displayed and execution waits for the user. This mechanism can be useful when events are infrequent and the requested action is easy to evaluate. It becomes less reliable as prompt volume grows.

Anthropic's Claude Code telemetry provides a recent real-world example. In March 2026, Anthropic reported that users approved approximately 93 percent of permission prompts. The company described approval fatigue as a practical problem and introduced an auto mode that uses classifiers to automate safer permission decisions while retaining stronger controls for riskier actions (Anthropic, 2026). The point is not that a 93 percent approval rate proves users made poor decisions; it does show that a control mechanism can become dominated by routine approval behavior at scale.

The broader human-AI literature explains why this matters. When users face repeated recommendations or explanations, trust, workload, verification difficulty, and cognitive effort influence whether they engage critically (Buçinca et al., 2021; Fok & Weld, 2024; Romeo & Conti, 2026). For agentic systems, approval prompts add a throughput dimension: the machine can create review opportunities faster than the human can evaluate them carefully.

Principle 1 (Exception-based oversight). Human review should be triggered by consequence-relevant boundary crossings, uncertainty, or anomalies rather than by the raw number of agent actions.

Principle 2 (Approval quality over approval count). Increasing the number of approval checkpoints does not necessarily increase meaningful control when the added checkpoints reduce the attention available to evaluate each one.

These principles favor control by policy. Instead of approving every harmless shell command, define a sandbox in which low-consequence commands are already authorized. Instead of approving every small purchase, define a spending envelope and escalate exceptions. Instead of requiring a human to inspect every routine message, constrain the class of messages the agent may send and route unusual content to review. This is not the removal of oversight; it relocates oversight from repeated micro-approvals to boundary design and exception handling.

6. Three Autonomy Bands

The framework can be operationalized as three broad autonomy bands. The bands are heuristics rather than validated thresholds, and organizations should define them according to domain-specific consequence tolerances.

Table 2. Heuristic autonomy bands for agent actions

Band	Typical profile	Primary control pattern	Illustrative actions
Pre-authorized autonomy	Low consequence, highly reversible, narrow authority	Policy boundary + logging	Read approved data; run tests; format code; generate drafts.
Policy-controlled autonomy	Moderate consequence or wider authority, but bounded and recoverable	Sandboxing, automated checks, asynchronous review, monitoring	Modify files in an isolated environment; create pull requests; change non-production configuration.

Band	Typical profile	Primary control pattern	Illustrative actions
Human escalation	High consequence, hard to reverse, broad authority, or unresolved ambiguity	Synchronous human decision before execution	Deploy to production; delete remote data; transfer substantial funds; change access permissions; publish sensitive information.

Reversibility is especially useful because it separates many superficially similar actions. Editing a local branch and modifying a production database may both be 'write' operations, but they have different recovery properties and blast radii. Authority provides another distinction: an agent that can draft a transfer request has different operational authority from an agent that can execute the transfer. Risk then captures the consequence if the action is wrong.

The bands should not be interpreted as permanent labels for tools. The same tool call can move between bands depending on context. Sending an internal test email and sending a legally consequential notice use the same basic capability but differ in consequence and delegated authority. A useful oversight system therefore routes actions using task and state context, not just tool names.

7. Four Modes of Oversight

Attention-constrained oversight also requires a broader temporal view. Human judgment can shape the system before, during, at exceptions, and after execution. Four modes are particularly useful: boundary setting, monitoring, escalation, and audit. These modes overlap with but are not identical to the empirical categories reported by Dhanorkar et al. (2026), whose developer interviews identify a priori control, co-planning, real-time monitoring, and post hoc review. The convergence is important because it suggests that practical oversight already extends beyond synchronous approval.

Table 3. Four complementary modes of oversight

Mode	When	Purpose	Examples
Boundary setting	Before execution	Define the autonomy envelope and limit possible failures	Permissions, budgets, tools, data access, prohibited actions, sandboxes.
Monitoring	During execution	Detect signals that the workflow is leaving the expected operating region	Unexpected spending, repeated retries, unusual tool calls, large changes, drift indicators.
Escalation	At exceptions	Concentrate human judgment on consequential or ambiguous decisions	Threshold crossings, low verifiability, policy conflicts, high-authority actions.
Audit	After execution	Detect drift, recurring failure modes, and weaknesses in policies	Sample review, failure analysis, near-miss review, trend monitoring, policy updates.

Boundary setting often provides the highest leverage because it determines what failures are possible before attention becomes necessary. Monitoring provides selective visibility without requiring the human to inspect every step. Escalation is the mode most closely associated with human-in-the-loop, but it should be reserved for decisions where human judgment has material value. Audit allows high-volume low-risk work to remain scalable while still producing evidence about drift, near misses, and control effectiveness.

A system that uses only escalation is fragile because every failure must be caught synchronously. A system that uses only audit can discover serious errors too late. A mature architecture combines all four modes, shifting the human role from repeated confirmation toward designing boundaries, interpreting exceptions, and learning from system behavior.

7.1 Integration with Delegation Contracts

Attention-constrained oversight addresses how human judgment is allocated during execution. The companion working paper *Delegation Contracts for Agentic AI* addresses the preceding question of what is being delegated by specifying objective, context, constraints, acceptance criteria, authority, and escalation. The two frameworks operate at different layers rather than competing for the same role.

The mapping is direct. Constraints and authority define the boundary-setting envelope; objective and context help interpret the consequence profile and monitoring signals; acceptance criteria support monitoring and audit; and escalation conditions define when control should return to a human. The delegation contract therefore specifies the task and its decision boundary, while attention-constrained oversight determines where finite human review capacity is spent as execution unfolds.

8. Design Implications for Agentic AI

First, human attention should be budgeted explicitly. Teams already budget compute, latency, API cost, and error rates; oversight capacity should be treated similarly. A workflow that requires a person to review hundreds of low-value events per hour should not be described as safely human-supervised merely because every event presents an approval button.

Second, escalation should carry decision-ready context. Fok and Weld's (2024) emphasis on verifiability suggests that an alert is useful only when the reviewer can evaluate the action efficiently. An escalation should therefore summarize the proposed action, relevant evidence, expected consequences, alternatives, and the reason the policy boundary was crossed. Dumping a long chain of reasoning into a prompt may increase cognitive burden without improving verification.

Third, containment can substitute for some synchronous approval without eliminating control. If an agent operates inside a sandbox, with constrained credentials, limited spending authority, or reversible state transitions, the system can safely tolerate a wider range of autonomous behavior. NIST's AI Risk Management Framework similarly emphasizes risk management across the lifecycle rather than relying on a single control point (Tabassi, 2023). The practical implication is that autonomy should be increased by reducing blast radius, not merely by trusting the model more.

Fourth, oversight metrics should evaluate decision quality rather than human touch count. Useful measures may include escalation precision, missed consequential events, reviewer reversal rates, time-to-decision, prompt volume per operator, post-hoc defect discovery, and the fraction of escalations that reviewers can independently verify. These measures are proposed as a research agenda rather than validated metrics in this paper.

Fifth, signs of degraded review conditions should change the oversight policy. Unusually rapid approvals, long runs of near-identical confirmations, rising prompt volume, or repeated post-approval reversals can be treated as signals that meaningful review may be deteriorating. These signals do not prove fatigue. They indicate that the system should not respond by creating still more synchronous prompts. A conservative response is to increase escalation selectivity, batch or defer low-consequence decisions, narrow high-authority operations, or temporarily pause actions whose consequence profile requires genuine human judgment.

Finally, the autonomy boundary should remain revisable. Actions that repeatedly prove safe and easy to recover may move toward policy-controlled autonomy; actions associated with near misses, growing

authority, cumulative exposure, or degraded review conditions may move toward stronger controls. The objective is not maximal autonomy or maximal supervision. It is a stable allocation in which machine execution remains bounded and human judgment remains scarce enough to matter.

9. Limitations and Research Agenda

The framework is conceptual and has several limitations. First, risk, reversibility, and authority are multidimensional and domain-specific. This paper specifies their direction of influence but does not provide validated numeric scales or universal thresholds. Empirical work is needed to test whether organizations can score these variables consistently and whether such scores predict better routing decisions.

Second, human attention is not a fixed quantity. Expertise, fatigue, time pressure, interface design, incentives, and task complexity can change how much meaningful review a person can perform. The attention-budget model is therefore an abstraction intended to prevent the unrealistic assumption of unlimited review capacity, not a claim that attention can be measured with a single stable number.

Third, the cumulative-exposure extension is intentionally simple. Summing exposure across actions can flag repeated activity, but real sequences may contain nonlinear interactions, threshold effects, path dependence, or correlated failures that a simple aggregate cannot represent. Future work should test richer sequence-level models without losing the framework's operational simplicity.

Fourth, the proposed autonomy bands and oversight modes require empirical comparison against alternatives. A minimal validation design could compare two matched workflows: an approval-by-event baseline in which each qualifying action requires synchronous confirmation, and a three-band policy-control condition using pre-authorized, policy-controlled, and human-escalation bands. Experiments could vary action consequence, prompt frequency, and reviewer workload while measuring missed consequential events, escalation precision, reviewer reversal rates, time-to-decision, review burden, and post-hoc defects. The key question is whether policy-based routing preserves or improves detection of consequential events while consuming less human attention. Field studies are also needed because laboratory tasks may understate interruptions, organizational pressures, and accountability structures present in real deployments.

10. Conclusion

Human oversight is often treated as a binary feature: either a person is in the loop or the AI is autonomous. Agentic systems make that framing increasingly inadequate. The meaningful design question is how much autonomy a particular action can tolerate and where scarce human judgment should be spent.

The framework developed here organizes that decision around risk, reversibility, authority, and human attention. Risk, reversibility, and authority define the consequence profile of an action; human attention constrains how much synchronous review can be meaningful. Cumulative exposure extends the analysis beyond isolated actions, while adaptive safeguards recognize that review conditions themselves can deteriorate. From that premise follow two practical conclusions. Oversight should be exception-based rather than uniformly event-based, and the quality of human approvals matters more than their count. Boundary setting, monitoring, escalation, and audit then provide complementary ways to maintain control across the workflow.

The central claim is simple: a human approval is not valuable because it is human. It is valuable when it preserves a real opportunity for judgment. Reliable agentic autonomy therefore requires systems that protect human attention as carefully as they protect permissions, credentials, and data.

AI Use Disclosure

OpenAI's ChatGPT was used only for literature discovery and bibliographic verification. The autonomy-and-oversight framework and underlying concepts originate in the author's prior published work (Kao, 2026). The author independently reviewed all cited sources and remains solely responsible for the final manuscript, its arguments, source selection, and citations.

References

Anthropic. (2026, March 25). How we built Claude Code auto mode: A safer way to skip permissions. Anthropic Engineering. https://www.anthropic.com/engineering/claude-code-auto-mode

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). 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). https://doi.org/10.1145/3449287

Dhanorkar, S., Passi, S., & Vorvoreanu, M. (2026). Human oversight of agentic systems in practice: Examining the oversight work, challenges, and heuristics of developers using software agents. Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency, 6438-6465. https://doi.org/10.1145/3805689.3812402

Fok, R., & Weld, D. S. (2024). In search of verifiability: Explanations rarely enable complementary performance in AI-advised decision making. AI Magazine, 45, 317-332. https://doi.org/10.1002/aaai.12182

Kao, J. (2026). How to stay in control when AI does the work: A practical framework for delegation, judgment, and accountability. Independent publication.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50-80. https://doi.org/10.1518/hfes.46.1.50_30392

Lu, Z., Wang, D., & Yin, M. (2024). Does more advice help? The effects of second opinions in AI-assisted decision making. Proceedings of the ACM on Human-Computer Interaction, 8(CSCW1), Article 217, 1-31. https://doi.org/10.1145/3653708

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2), 230-253. https://doi.org/10.1518/001872097778543886

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). 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. https://doi.org/10.1109/3468.844354

Passi, S., Dhanorkar, S., & Vorvoreanu, M. (2024). Appropriate reliance on generative AI: Research synthesis (MSR-TR-2024-7). Microsoft Research.

Romeo, G., & Conti, D. (2026). Exploring automation bias in human-AI collaboration: A review and implications for explainable AI. AI & Society, 41, 259-278. https://doi.org/10.1007/s00146-025-02422-7

Schemmer, M., Kuehl, N., Benz, C., Bartos, A., & Satzger, G. (2023). Appropriate reliance on AI advice: Conceptualization and the effect of explanations. Proceedings of the 28th International Conference on Intelligent User Interfaces, 410-422. https://doi.org/10.1145/3581641.3584066

Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

Wittens, M. M. J., Kacafírková, K. S., Elprama, S. A., & Jacobs, A. (2026). The challenge of designing for meaningful effective human oversight in rapidly evolving complex AI systems. HHAI 2026: Proceedings of the 5th International Conference on Hybrid Human-Artificial Intelligence. https://doi.org/10.3233/FAIA260543

Zhu, L., Lu, Q., Ding, M., Lee, S. U., & Wang, C. (2026). Designing meaningful human oversight in AI. AI and Ethics, 6, Article 286. https://doi.org/10.1007/s43681-026-01147-7