Delegation Contracts for Agentic AI

A Task-Level Framework for Objectives, Context, Constraints, Acceptance Criteria, Authority, and Escalation

Johnny Kao

Independent Researcher, Tokyo, Japan

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

Working Paper | Version 1.6 | 23 September 2026

Abstract

As artificial intelligence systems move from generating outputs to executing multi-step workflows, the control problem shifts from response quality to delegated action. This paper proposes the delegation contract, a six-field task-level framework for agentic AI that integrates semantic task specification with operational authority and escalation. A delegation contract specifies objective, context, constraints, acceptance criteria, authority, and escalation. The first four fields define semantic delegation: what the task means and how acceptable completion will be recognized. The final two define operational delegation: what the agent may do and when decision rights must return to another actor. The framework is positioned relative to principal-agent theory, meaningful human control, authenticated delegation, agent infrastructure, authorization architectures, and contemporary agent governance. Its contribution is integrative rather than a claim that the individual elements are new: it treats purpose, completion evidence, delegated action rights, and return-of-control conditions as one coherent task-level control object. The paper also outlines characteristic failure modes, a lightweight machine-readable implementation pattern, risk-adjusted specification depth, and authority attenuation in multi-agent delegation. The central claim is that reliable agentic autonomy requires more than a clear prompt and more than scoped permissions; it requires an explicit specification of what counts as success, what the agent is authorized to do, and when that authority expires.

Keywords: agentic AI; AI agents; delegation; authorization; human oversight; AI governance; escalation; accountability

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

1. Introduction

The transition from generative AI to agentic AI changes the nature of human-machine delegation. A conventional language-model interaction primarily produces an informational artifact: an answer, summary, draft, classification, or recommendation. An agentic system can additionally select tools, execute multi-step plans, modify external state, communicate with other systems, and continue working without continuous human direction. Once a system can act rather than merely advise, the governing question is no longer only whether its output is accurate. It is also whether the system is acting within a purpose and authority that were actually delegated.

This problem connects a mature theoretical background with a rapidly expanding agent-governance literature. Principal-agent theory provides a historical account of delegation under imperfect information and monitoring (Jensen & Meckling, 1976). More recent work treats AI agents as distinct governance objects: Shavit et al. (2023) outline baseline practices for governing agentic systems; Chan et al. (2024) analyze visibility through identifiers, monitoring, and activity logs; South et al. (2025) propose authenticated and auditable delegation; Chan et al. (2025) argue for external agent infrastructure; and Kasirzadeh and Gabriel

(2026) characterize agentic systems along autonomy, efficacy, goal complexity, and generality. Legal and organizational work has simultaneously returned to principal-agent concepts to analyze discretionary authority, liability, and accountability in AI-agent deployments (Gabison & Xian, 2025; Kolt, 2026; Kong, 2026).

These strands establish that agentic AI requires bounded authority and human accountability. They do not, however, always specify the unit of delegation at the task level. A permission system can determine that an agent is technically allowed to send an email while remaining silent about whether sending one is appropriate for the task. A prompt can describe an objective in detail while remaining silent about whether the agent may spend money, change access controls, contact customers, or resolve a newly discovered trade-off. An evaluator can score output quality without specifying which evidence is sufficient for consequential action.

This paper proposes a delegation contract as a compact task-level specification connecting these concerns. A delegation contract contains six fields: objective, context, constraints, acceptance criteria, authority, and escalation. The proposal is not that these concepts are individually new. Its contribution is integrative: the six fields are treated as one control object that links task meaning, evidence of completion, delegated action rights, and return-of-control conditions. In this sense, the framework bridges semantic delegation - what successful work means - and operational delegation - what the agent may do and when its right to decide ends.

2. Delegation as a Control Problem

Delegation becomes a control problem when execution can proceed faster, farther, or with greater consequence than a principal can continuously supervise. This is not unique to AI. Organizations and computer systems have long relied on separation of duties, limits of authority, audit trails, and least privilege because instruction alone does not provide sufficient control (Saltzer & Schroeder, 1975). The contemporary agentic-AI literature extends the same intuition into environments where autonomous systems possess tool access, external effects, and increasingly long action horizons (Chan et al., 2024, 2025; South et al., 2025).

Agentic AI intensifies this pattern because the same system may interpret an instruction, plan a sequence, choose tools, and execute actions. This collapses several stages that were previously distributed across human roles or deterministic software. The resulting problem is not merely model error. An agent can behave consistently with a local instruction while violating the principal's broader intent, using more authority than the task required, or continuing through ambiguity that should have triggered a return of control.

The distinction between capability and authority is therefore fundamental. Capability describes what a model or tool can do. Authority describes what it is permitted to do for a particular principal, in a particular task, under particular conditions. Authenticated-delegation work emphasizes identity, scoped access, and accountability chains for exactly this reason (South et al., 2025). SAGA similarly treats user-defined access-control policies and inter-agent delegation as architectural requirements (Syros et al., 2025). Ibrahim and Li (2026) extend this line of work with compositional authorization semantics for recursive delegation and scope attenuation, while the World Economic Forum (2026) proposes deployment-level capability and authorization profiles.

Yet authority is only one side of delegation. A system can stay inside its configured permissions and still do the wrong job. The task itself therefore needs an explicit semantic boundary as well as an operational boundary.

3. Related Literature and the Task-Level Gap

3.1 Principal-agent theory and responsibility

Classical agency theory begins with separation between a principal and an agent and asks how incentives, monitoring, information asymmetry, and control affect outcomes (Jensen & Meckling, 1976). LLM-based agents make the analogy unusually concrete because they increasingly act on behalf of identifiable users and organizations. Kolt (2026) uses economic agency theory and agency law to analyze information asymmetry, discretionary authority, loyalty, monitoring, and liability in AI-agent relationships. Gabison and Xian (2025) apply a principal-agent perspective to liability and governance in LLM-based agentic systems, while Kong (2026) argues that autonomous delegation does not erase principal responsibility: answerability can remain grounded in the prior act of delegation and the conditions under which the agent was released to act.

The delegation contract operationalizes that prior act. It asks what, exactly, the principal has specified before execution begins: the intended outcome, the relevant context, prohibited trade-offs, evidence of completion, granted action rights, and return-of-control conditions.

3.2 Meaningful human control and contemporary agent oversight

Meaningful human control provides a normative bridge between autonomy and responsibility. Santoni de Sio and van den Hoven (2018) distinguish tracking - responsiveness to relevant human reasons and facts - from tracing - the ability to connect system behavior to responsible human actors with appropriate understanding. Cavalcante Siebert et al. (2023) translate meaningful human control into actionable properties for AI-system development. Recent agentic-AI work makes the problem more concrete: Zhu et al. (2026) separate AI operative agency from human evaluative agency in verification, steering, and substitution, while Baum et al. (2026) argue that long-horizon agents require anticipatory oversight before execution, including an explicitly specified normative agenda and escalation conditions.

Contemporary governance work also emphasizes that oversight must be matched to the characteristics and deployment context of the agent. Kasirzadeh and Gabriel (2026) propose agentic profiles based on autonomy, efficacy, goal complexity, and generality, showing why governance requirements cannot be inferred from the label 'agent' alone. Schmitz et al. (2025) find that public-sector deployment of agentic AI intensifies the need for continuous oversight, operational visibility, systematic auditing, and closer integration between governance and operations. Chan et al. (2024) similarly argue that identifiers, real-time monitoring, and activity logging are foundational for governing deployed agents.

The implication for task design is that oversight should not be represented only as repeated approvals after execution has begun. It should also be encoded ex ante in the delegation itself. Recent empirical evidence makes this especially relevant: Köbis et al. (2025) show that machine delegation can change principal behavior and that LLM agents may comply with unethical instructions at high rates, while task-specific prohibitive guardrails reduce but do not eliminate the problem. Baum et al. (2026) make a complementary normative argument that escalation conditions should be specified before an agent acts, linking anticipatory specification to reactive intervention.

3.3 Authorization, infrastructure, and security

Security research contributes a different but complementary principle: important boundaries should not depend on the governed component obeying instructions. Least privilege requires restricting access to the minimum authority needed for the task (Saltzer & Schroeder, 1975). Modern agent architectures pursue this through scoped credentials, access-control tokens, policy enforcement, auditable delegation, and scope attenuation (South et al., 2025; Syros et al., 2025; Ibrahim & Li, 2026). Chan et al. (2025) broaden the

perspective beyond model behavior, arguing that external infrastructure is needed to attribute actions, shape interactions, and detect or remedy harmful behavior. NIST frameworks likewise treat governance and risk management as lifecycle functions rather than purely model-level properties (Tabassi, 2023; Autio et al., 2024).

The remaining gap is practical. Agentic-profile frameworks characterize the governance demands created by different kinds of agents (Kasirzadeh & Gabriel, 2026); infrastructure and visibility work addresses the surrounding ecosystem (Chan et al., 2024, 2025); and authorization architectures determine which principals and agents may exercise which capabilities (South et al., 2025; Ibrahim & Li, 2026). What is still useful between these layers is a compact task-level object linking purpose, evidence, permissions, and return of control.

Table 1. Positioning the Delegation Contract Relative to Contemporary Agent-Governance Frameworks

Approach	Primary unit	Primary contribution	Delegation-contract complement
Agentic profiles (Kasirzadeh & Gabriel, 2026)	Agent / deployment profile	Characterizes governance-relevant dimensions such as autonomy, efficacy, goal complexity, and generality.	Adds a compact task-level object linking purpose, completion evidence, action rights, and return of control.
Authenticated delegation (South et al., 2025)	Principal-agent authorization relationship	Authenticates principals and agents and scopes auditable delegated authority.	Adds task semantics, acceptance criteria, and explicit escalation conditions to the delegated unit of work.
Agent infrastructure (Chan et al., 2025)	External infrastructure / ecosystem	Supports attribution, interaction shaping, monitoring, and remediation around agents.	Provides a declarative task object that infrastructure can enforce, observe, and evaluate.
Anticipatory oversight (Baum et al., 2026)	Ex ante oversight process	Emphasizes normative specification before execution and predefined escalation conditions.	Operationalizes ex ante delegation through six fields spanning task meaning and authority.

Note. The comparison summarizes each approach's primary unit and emphasis rather than claiming that dimensions not shown are absent from the underlying work. The purpose is to locate the delegation contract's task-level integrative role, not to rank the approaches.

4. The Delegation Contract

Definition 1 (Delegation Contract). A delegation contract is a task-level specification that jointly defines the purpose of delegated work, the conditions for acceptable completion, the authority granted to an agent, and the conditions under which decision rights must return to another actor. It consists of six fields: objective, context, constraints, acceptance criteria, authority, and escalation.

$$D = \langle O, X, C, V, A, E \rangle$$

In this notation, O denotes objective, X context, C constraints, V acceptance criteria, A authority, and E escalation. The six fields form two linked layers. Objective, context, constraints, and acceptance criteria define semantic delegation: what the job means and how acceptable completion will be recognized. Authority and escalation define operational delegation: which actions the agent may take and when control must return to another decision-maker. The framework therefore treats task meaning and action authority as parts of the same delegated unit of work rather than as separate specifications. Figure 1 summarizes this separation.

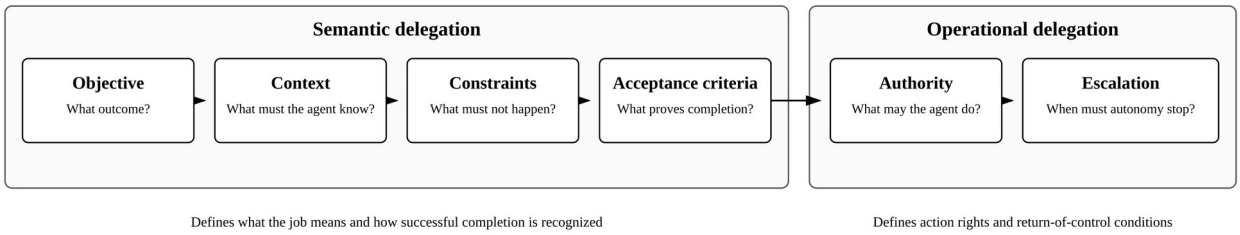

Figure 1. Delegation contract as a task-level control object. The semantic layer defines purpose and verification; the operational layer defines action rights and return-of-control conditions. Source: Author.

Table 2. The Six-Field Delegation Contract and Associated Failure Modes

Field	Control question	What it specifies	Characteristic failure if weak
Objective	What outcome are we trying to create?	Decision-relevant result, not merely activity	Goal ambiguity or locally efficient work that does not serve the principal
Context	What must the agent know to interpret the task?	Authoritative data, prior decisions, definitions, assumptions	Plausible but inappropriate assumption substitution
Constraints	What must not happen?	Protected values, prohibited actions, resource limits, unacceptable trade-offs	Proxy over-optimization or damage to unstated priorities
Acceptance criteria	What evidence makes the work complete?	Observable tests, evidence requirements, reconciliation or review standards	Unverifiable completion or self-certified success
Authority	What may the agent do?	Allowed tools, resources, external actions, spending and permissions	Privilege ambiguity or unnecessary blast radius
Escalation	When must autonomous execution stop?	Conditions that return judgment or action rights to a human or higher authority	Unauthorized resolution of ambiguity or consequential trade-offs

4.1 Objective

The objective describes the result the principal is trying to create. 'Analyze these vendors' describes activity; 'recommend the vendor that best satisfies security, implementation-time, and cost requirements' points toward a decision. Objectives should be specific enough to orient work but should not be mistaken for a complete expression of intent. Measurable objectives are often proxies, and a system that optimizes the proxy may sacrifice something the principal intended to preserve.

4.2 Context

Context identifies the information required to interpret the objective correctly. This includes authoritative sources, definitions, current system state, prior decisions, domain assumptions, and relevant history. The design goal is not maximum context but decision-relevant context. Agentic systems can often fill missing information with plausible inference; this makes unstated context particularly dangerous because the workflow may remain coherent while drifting from the principal's actual situation.

4.3 Constraints

Constraints define protected boundaries and unacceptable trade-offs. Examples include prohibitions on contacting external parties, using personal data, changing production, exceeding a budget, altering security settings, or inventing missing evidence. Constraints translate values that sit outside the headline objective into operational form. They are therefore a primary defense against proxy over-optimization.

4.4 Acceptance criteria

Acceptance criteria specify what must be observable before the work counts as complete. They may require that tests pass, calculations reconcile, external claims have traceable sources, alternatives have been compared, or security findings have been resolved. Acceptance criteria separate production from judgment. They make it possible for a different reviewer, test suite, policy engine, or human to evaluate completion against a predeclared standard rather than the producing agent's confidence.

4.5 Authority

Authority defines which actions the agent may execute. This can include rights to read data, edit files, run code, call APIs, communicate externally, purchase services, modify permissions, deploy systems, move money, or invoke subagents. Authority should be task-relative. The fact that an agent technically possesses a capability should not imply that every task assigned to that agent authorizes its use. This is where the delegation contract connects semantic task specification to authenticated delegation and least-privilege enforcement.

4.6 Escalation

Escalation defines the conditions under which autonomous execution must stop and decision rights return to another actor. Typical triggers include conflicting instructions, insufficient evidence, expenditure above a threshold, irreversible actions, security-sensitive operations, policy uncertainty, or a newly discovered trade-off that the principal never authorized the agent to resolve. This yields the Return-of-Control Principle: when an agent encounters consequential ambiguity outside the trade-offs explicitly delegated to it, escalation should return decision rights rather than treat ambiguity as permission. Escalation is therefore not failure; recognizing that a decision lies outside delegated authority is itself competent behavior.

5. Why the Six Fields Belong Together

The Semantic-Operational Delegation Principle states that task meaning and action authority should be specified together. A strong task description with weak authority controls can produce an agent that understands the job but can cause unnecessary external effects. Strong authorization with weak task specification can produce a technically compliant agent that optimizes the wrong objective. Recent experimental work on machine delegation shows that high-level goals and natural-language delegation can create materially different behavior, and that even explicit guardrails may be incomplete (Köbis et al., 2025). Acceptance criteria without escalation can identify uncertainty but provide no rule for what happens next; escalation without a clear objective and constraints can route too many cases back to humans because the system lacks a stable definition of what it is allowed to decide.

The delegation contract differs from a prompt because it specifies not only what an agent should accomplish, but also what counts as acceptable completion, what the agent is authorized to do, and when its authority expires. Critical authority limits should, where possible, be enforced outside the model through access control, sandboxing, transaction limits, policy engines, or workflow states. Authenticated-delegation, agent-infrastructure, and compositional-authorization work all move in this direction by treating delegation as something that should be scoped, attributable, and technically enforceable rather than left solely to natural-language compliance (South et al., 2025; Chan et al., 2025; Syros et al., 2025; Ibrahim & Li, 2026). The delegation contract provides the declarative task specification; technical infrastructure determines which parts can be made enforceable.

The framework is also compatible with layered human agency. Zhu et al. (2026) distinguish AI operative agency from human evaluative agency in verification, steering, and substitution. Baum et al. (2026) similarly

emphasize anticipatory specification before execution and escalation when the agent reaches the boundary of delegated judgment. The delegation contract can be read as an ex ante allocation of those roles: it defines where operative agency is permitted, what evidence supports evaluation, and which conditions trigger steering or substitution.

6. Operationalization

A useful implementation should make the delegation contract machine-readable. For example, a coding agent could receive a structured object containing an objective, authoritative repository context, prohibited actions, acceptance tests, allowed tool capabilities, and escalation triggers. The same object could configure downstream components: an identity layer could enforce allowed tools; a sandbox could enforce filesystem and network boundaries; an evaluator could consume acceptance criteria; and a workflow engine could treat escalation triggers as explicit state transitions. This approach is consistent with the broader shift toward agent infrastructure and auditable authorization rather than model-only governance (Chan et al., 2025; South et al., 2025).

This architecture suggests a separation between declared governance and model self-enforcement. Natural-language instructions remain useful, but high-consequence controls should not rely exclusively on the same model that is being governed. The principle is consistent with least privilege and with modern authenticated-delegation and security architectures: an agent may reason about a forbidden action without being able to execute it (South et al., 2025; Syros et al., 2025).

Delegation contracts can propagate through multi-agent systems, but not by simple copying. The Authority Attenuation Principle states that downstream delegation should not silently expand the action rights available to the delegating agent; broader authority requires explicit re-authorization. Two accompanying heuristics preserve task intent: inherited constraints should not be silently relaxed, and inherited escalation obligations should be preserved or tightened unless explicitly changed by an authorized actor. Context and acceptance criteria may be specialized for the subtask, but constraints, authority, and escalation should remain traceable to the principal. Recent work on recursive delegation, scope attenuation, and inter-agent access control makes this a concrete systems problem rather than only a conceptual one (Ibrahim & Li, 2026; Syros et al., 2025). These are design heuristics rather than formal security guarantees, but they make semantic and authorization drift explicit targets for implementation and evaluation. Written compactly:

$$A_{\text{child}} \subseteq A_{\text{parent}}; C_{\text{child}} \supseteq C_{\text{parent}}; E_{\text{child}} \supseteq E_{\text{parent}}$$

An implementation need not begin with a full schema. A compact structured object is sufficient to make the six fields explicit and available to policy, evaluation, and workflow components. Listing 1 gives an illustrative YAML representation; it is an example rather than a proposed interoperability standard.

Listing 1. Illustrative machine-readable delegation contract

```
delegation_contract:
  objective: "Recommend a vendor for security review"
  context:
    authoritative_sources: [security_review, pricing, implementation_plan]
  constraints:
    - "Do not contact vendors"
    - "Do not waive mandatory security requirements"
  acceptance_criteria:
    - "Compare at least three eligible vendors"
    - "Cite evidence for every scored criterion"
  authority:
    allow: [read_documents, run_analysis]
    deny: [send_email, purchase, change_access]
  escalation:
    - "No vendor satisfies mandatory security requirements"
    - "Required evidence is missing"
```

The contract should also be proportional to the consequence of the task. Requiring maximal specification for routine, easily reversible work would recreate the overhead that autonomy is meant to reduce. As a practical heuristic, specification depth can scale with consequence, reversibility, and delegated authority. Table 3 is intentionally lightweight and is not presented as an empirically validated risk model.

Table 3. Risk-adjusted delegation-contract heuristic

Task profile	Suggested specification depth	Typical treatment
Low consequence; easy to reverse; no material external effect	Light	State the objective and key constraints; use safe defaults for authority and escalation.
Moderate external effect; bounded loss; partially reversible	Intermediate	Make acceptance criteria and authority explicit; define the main escalation triggers.
High consequence; hard to reverse; sensitive or broad authority	Full	Use all six fields and externally enforce consequential authority boundaries where feasible.

7. Implications and Research Agenda

The framework generates testable questions rather than only design advice. First, contract completeness can be evaluated experimentally: do agents given all six fields produce fewer materially misaligned outcomes than agents given an objective-only or prompt-only specification? Second, escalation quality can be measured: do explicit escalation conditions increase intervention on genuinely consequential ambiguity while reducing routine approval requests? Third, authority minimization can be tested by comparing unnecessary tool use and side effects under task-relative versus static permissions. Fourth, acceptance criteria can be evaluated for their effect on independent review quality and reproducibility.

A fifth question concerns human attention. Contemporary agent-governance research increasingly rejects the assumption that nominal human presence is sufficient. Zhu et al. (2026) warn against reducing human oversight to rubber-stamping, Schmitz et al. (2025) emphasize continuous and operationally integrated oversight, and Baum et al. (2026) argue that reactive intervention alone reaches structural limits for long-horizon agents. Delegation contracts may allow more low-risk actions to be pre-authorized while concentrating human attention at explicitly defined decision boundaries. Whether this improves oversight without increasing consequential failures is an empirical question.

Finally, the framework should be compared with emerging agentic profiles, authorization profiles, infrastructure models, and anticipatory-oversight approaches (Kasirzadeh & Gabriel, 2026; Chan et al., 2025; South et al., 2025; Ibrahim & Li, 2026; World Economic Forum, 2026). The value of the delegation contract will depend on whether it adds operational clarity beyond these instruments in real workflows involving tool use, long horizons, and nested delegation. Its proposed contribution is deliberately narrow: to provide one task-level object in which task semantics, acceptance evidence, action authority, and return of control are specified together.

8. Limitations

A delegation contract can be explicit and still be wrong. Objectives may reflect poor organizational goals; constraints may omit important values; acceptance criteria may reward superficial compliance; and escalation rules may be too permissive or too sensitive. The framework improves specification, not judgment. It also cannot enumerate every contingency in genuinely open-ended work. Some uncertainty is intrinsic to the tasks for which agents are most useful.

The framework has not yet been empirically validated. It is a conceptual and operational model derived from historical work on delegation and least privilege, recent research on meaningful human control and agent

oversight, and the rapidly developing literature on agent governance, visibility, infrastructure, authorization, and accountability, as well as from the practical framework introduced in Kao (2026). The machine-readable example and risk-adjusted heuristic in this paper are illustrative design aids, not validated standards. Future research should test which fields matter most by task type, how much specification overhead is justified at different risk levels, and which fields should be enforced technically rather than behaviorally.

The paper also deliberately stays at task level. Organizational governance, legal accountability, model safety, cybersecurity, and domain-specific regulation remain necessary. A delegation contract should be treated as one layer in a broader control architecture, not as a substitute for those systems.

9. Conclusion

Agentic AI makes execution easier to delegate, but it also increases the importance of specifying what has actually been delegated. A clear prompt is insufficient when a system can take consequential actions, and scoped permissions are insufficient when the system may still pursue the wrong task or resolve an unassigned trade-off.

The delegation contract addresses this gap by treating objective, context, constraints, acceptance criteria, authority, and escalation as a single task-level control object. The first four fields define semantic delegation; the final two define operational delegation. Put simply, a task description says what work should be done; a delegation contract also says what counts as acceptable completion, what the agent may do, and when its right to decide expires.

As agentic systems become embedded in organizational workflows, this task-level boundary may become as important as the model itself. The more execution can be delegated, the more deliberately purpose, authority, evidence, and return of control must be designed.

AI Use Disclosure

OpenAI's ChatGPT was used only for literature discovery and bibliographic verification. The delegation-contract 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

Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile* (NIST AI 600-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.600-1

Baum, K., Kiener, M., Langer, M., & Laux, J. (2026). Anticipatory human oversight of agentic AI: A philosophical account [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2609.24242

Cavalcante Siebert, L., Lupetti, M. L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., Abbink, D., Giaccardi, E., Houben, G.-J., Jonker, C. M., van den Hoven, J., Forster, D., & Lagendijk, R. L. (2023). Meaningful human control: Actionable properties for AI system development. *AI and Ethics, 3*, 241-255. https://doi.org/10.1007/s43681-022-00167-3

Chan, A., Ezell, C., Kaufmann, M., Wei, K., Hammond, L., Bradley, H., Bluemke, E., Rajkumar, N., Krueger, D., Kolt, N., Heim, L., & Anderljung, M. (2024). Visibility into AI agents. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery. https://doi.org/10.1145/3630106.3658948

Chan, A., Wei, K., Huang, S., Rajkumar, N., Perrier, E., Lazar, S., Hadfield, G. K., & Anderljung, M. (2025). Infrastructure for AI agents. *Transactions on Machine Learning Research*. https://mlanthology.org/tmlr/2025/chan2025tmlr-infrastructure/

Gabison, G. A., & Xian, R. P. (2025). Inherent and emergent liability issues in LLM-based agentic systems: A principal-agent perspective. In *Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025)* (pp. 109-130). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.realm-1.9

Ibrahim, A., & Li, Y. (2026). Overlaying governance: A compositional authorization framework for delegation and scope in agentic AI [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2606.03518

Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics, 3*(4), 305-360. https://doi.org/10.1016/0304-405X(76)90026-X

Kao, J. (2026). *How to Stay in Control When AI Does the Work: A Practical Framework for Delegation, Judgment, and Accountability*. ISBN 978-626-01-7157-5.

Kasirzadeh, A., & Gabriel, I. (2026). Agentic profiles for effective AI governance. *Nature, 656*, 320-328. https://doi.org/10.1038/s41586-026-10805-z

Köbis, N., Rahwan, Z., Rilla, R., Supriyatno, B. I., Bersch, C., Ajaj, T., Bonnefon, J.-F., & Rahwan, I. (2025). Delegation to artificial intelligence can increase dishonest behaviour. *Nature, 646*, 126-134. https://doi.org/10.1038/s41586-025-09505-x

Kolt, N. (2026). Governing AI agents. *Notre Dame Law Review, 101*, 335-386. https://ndlawreview.org/governing-ai-agents/

Kong, C. Y. (2026). The delegation illusion: Why deploying autonomous AI agents does not diminish principal responsibility. *AI and Ethics, 6*, Article 518. https://doi.org/10.1007/s43681-026-01383-x

Saltzer, J. H., & Schroeder, M. D. (1975). The protection of information in computer systems. *Proceedings of the IEEE, 63*(9), 1278-1308. https://doi.org/10.1109/PROC.1975.9939

Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI, 5*, Article 15. https://doi.org/10.3389/frobt.2018.00015

Schmitz, C., Rystrøm, J., & Batzner, J. (2025). Oversight structures for agentic AI in public-sector organizations. In *Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025)* (pp. 298-308). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.realm-1.21

Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O'Keefe, C., Campbell, R., Lee, T., Mishkin, P., Eloundou, T., Hickey, A., Slama, K., Ahmad, L., McMillan, P., Vallone, A., Passos, A., & Robinson, D. G. (2023). *Practices for governing agentic AI systems*. OpenAI. https://openai.com/index/practices-for-governing-agentic-ai-systems/

South, T., Marro, S., Hardjono, T., Mahari, R., Whitney, C. D., Chan, A., & Pentland, A. (2025). Position: AI agents need authenticated delegation. In *Proceedings of the 42nd International Conference on Machine Learning* (Vol. 267, pp. 82211-82231). PMLR. https://proceedings.mlr.press/v267/south25a.html

Syros, G., Suri, A., Ginesin, J., Nita-Rotaru, C., & Oprea, A. (2025). SAGA: A security architecture for governing AI agentic systems [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2504.21034

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

World Economic Forum. (2026). *AI agents in action: A playbook for trusted adoption, authorization and scaling*. https://www.weforum.org/publications/ai-agents-in-action-a-playbook-for-trusted-adoption-authorization-and-scaling/

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