The Principal–AI Agent Problem: Governance in the Age of AI Agents and Principal–AI Agent Theory

Seiryu Ando

Graduate School of Economics, Kyoto University

ando.seiryu.48a@st.kyoto-u.ac.jp

November 1, 2026

Abstract

Across various areas of business and society, tasks and decisions traditionally undertaken by humans are increasingly being delegated to AI agents. This expansion of delegation raises the question of how to govern AI agents so that their behavior serves the objectives of those who delegate to them. Building on the conventional principal–agent problem and agency theory, this paper defines the principal–AI agent problem and presents a framework of "AI agency theory" that systematizes its underlying mechanisms, cost structure, and governance instruments.

The paper first reviews conflicts of interest, information asymmetry, agency costs, and governance through incentives, monitoring, and the allocation of authority in conventional theory. It then takes as its baseline analytical object an AI that follows externally supplied objectives and does not pursue private interests of its own. Even in this case, delegation can fail because the principal cannot adequately communicate objectives, constraints, and priorities, or sufficiently observe, evaluate, and control the AI's behavior. This paper calls the divergence between the principal's intentions and the AI's behavior arising from these constraints on delegation "AI agency slack." It defines AI agency costs as the governance costs of preventing, detecting, and correcting such divergence, together with the residual loss that remains after these measures. The corresponding governance instruments include the design and updating of instructions, monitoring, the specification of the scope of delegation, and the allocation of authority to approve, modify, and stop actions.

The paper next considers a relationship in which shareholders are the principal and an AI agent acting as a manager is the agent. Through comparison with corporate governance premised on human managers, it examines the arrangements for oversight and control that might be considered when management is delegated to an AI agent.

The discussion then extends to four relationships: managers and AI agents acting as workers; politicians and AI agents acting as bureaucrats; clients and AI agents acting as professionals such as lawyers and physicians; and limited partner (LP) investors and AI fund managers. For each relationship, the paper anticipates how conventional agency problems may change when human agents are replaced by AI, and what delegation losses and governance challenges may arise. The paper contributes by clarifying which elements of conventional agency theory should be retained and which should be reformulated for delegation to AI, thereby providing an analytical foundation for AI agency theory to understand diverse delegation relationships in firms and beyond.

Keywords: principal–AI agent problem; AI agency theory; AI agency costs; monitoring; delegation of authority; corporate governance

JEL classification: D82, D86, G34

1 Introduction

Across various areas of business and society, tasks and decisions traditionally undertaken by humans are increasingly being delegated to AI agents. Dots, announced by OpenAI in September 2026, continuously carries out work toward a user's goals using connected applications and a computer in the cloud. The company reports a case in which Dots identified unbilled work, prepared an invoice, and sent it after obtaining the user's approval (OpenAI, 2026). Muse, announced by Meta in the same month, is also offered as an agent that develops plans from a user's goals and proceeds with tasks such as booking travel and purchasing goods. It includes a mechanism for seeking the user's approval for actions such as sending emails and making purchases (Meta, 2026).

These services can be understood as delegation relationships that combine the communication of objectives, the granting of a degree of discretion, and human approval. When tasks are entrusted to AI, the questions are how to communicate the delegator's intentions, how much judgment to delegate, and how to oversee and modify its behavior. The expansion of delegation to AI heightens the need for a theoretical examination of these governance arrangements.

Relationships in which tasks or decisions are entrusted to others are a central object of analysis in principal–agent theory. Ross (1973) formalized the relationship in which an agent acts on behalf of a principal as a general economic problem. Jensen and Meckling (1976) defined an agency relationship as involving the delegation of decision-making authority and classified the costs arising when the interests of the principal and agent diverge into monitoring expenditures, bonding expenditures, and residual loss. These studies provide a foundation for analyzing the benefits of delegation and the costs of aligning an agent's behavior with the principal's interests.

When the agent is AI, however, the sources of the problem and the operation of governance instruments require renewed examination. The baseline analytical object in this paper is an AI agent that follows externally supplied objectives and does not pursue private interests of its own. Under this assumption, problems arising from human agents prioritizing their own interests cannot simply be applied to AI. On the other hand, not assuming the pursuit of private interests does not mean that the principal's intentions are fully realized. If the principal cannot adequately specify the desired objectives and constraints, undesirable outcomes may arise even when the AI faithfully follows the objec-

tives it has been given. Moreover, the principal may be unable to address the divergence if they cannot evaluate the information used or the actions selected by the AI, or if they cannot modify or stop execution even after identifying a problem.

Research connecting delegation to AI with existing theory is already developing. Humberd and Latham (2026) examine the evolution of AI within firms through the lens of agency theory and discuss monitoring and incentives at different stages of capability development. Hadfield-Menell and Hadfield (2019) connect the mismatch between an AI's reward function and the objective intended by humans to the problem of incomplete contracts, highlighting the role of institutions and norms in filling unspecified gaps. Furthermore, Ante (2026) focuses on delegation losses that remain even in the absence of conflicts of interest and introduces the concept of specification cost. These studies point to the need to distinguish problems of motivation from problems of specifying, verifying, and controlling objectives when considering delegation to AI.

Building on these insights, this paper presents a framework of AI agency theory for analyzing the principal–AI agent problem and its governance. A principal–AI agent relationship is defined here as a relationship in which a human or organization delegates tasks and a degree of decision-making and execution authority to an AI agent for its own purposes. The unit of analysis is not the performance of an individual model, but a delegation relationship comprising the principal's objectives, the instructions given to the AI, the authority delegated, and the arrangements for oversight and intervention. The central questions are how divergence between the principal's intentions and the AI's behavior arises and what governance is needed to address it.

This paper calls divergence between the principal's intentions and the AI's behavior arising from constraints on delegation "AI agency slack." In analyzing it, the paper distinguishes between leaving parts of instructions unspecified, specifying objectives that differ from the principal's intentions, and failing to comply with specified conditions. It also examines losses attributable to the delegation arrangement without equating these with general deficiencies in capability or bad outcomes due to chance. Incomplete specification of objectives may cause no loss if the gaps are appropriately filled, whereas even detailed instructions may cause losses if they misrepresent the principal's intentions. This distinction is necessary for considering governance instruments that address the source of the problem.

The paper further defines AI agency costs as the governance costs of preventing, detecting, and correcting AI agency slack, together with the residual loss that remains after governance measures are taken. Governance costs here include the costs of preparing and updating instructions, observing and verifying actions and outcomes, granting approval, making corrections, restricting authority, and stopping execution. Accordingly, stronger oversight and restrictions on authority are not regarded as desirable in themselves. While they may reduce losses, they also require human attention and judgment and may diminish the benefits of delegation to AI. The issue to be examined is which arrangements are desirable in light of the relationship between the benefits of delegation, governance costs, and residual loss.

Monitoring in this paper is understood as a process of observing the AI's use of information, behavior, and outcomes and connecting that information to revisions of instructions or interventions in behavior. In this baseline setting, self-restraint by a self-interested actor aware of being monitored is not the central channel through which governance operates. Rather, what matters is what can be observed and who can use that information to change what. This point also connects with Aghion and Tirole (1997), who distinguish formal decision rights from effective control supported by information. Giving a human the authority to approve or stop actions does not necessarily make that authority effective if the information needed for judgment or the time to intervene is lacking. Arrangements for understanding the AI's behavior must therefore be considered together with arrangements for changing it.

As a main application of this framework, the paper considers a relationship in which shareholders are the principal and an AI agent acting as a manager is the agent. An AI manager here refers to AI to which managerial decisions or business operations have been delegated. Through comparison with corporate governance premised on human managers, the paper examines possible arrangements for oversight and control when AI undertakes management. The basic issues are how to reflect shareholders' intentions in managerial objectives, how to evaluate the AI's decisions and outcomes, and how to modify or stop its actions when necessary.

The principal–AI agent problem is not limited to the relationship between shareholders and managers. This paper anticipates problems that may arise in four relationships: managers and AI agents acting as workers; politicians and AI agents acting as bureau-

crats; clients and AI agents acting as professionals such as lawyers and physicians; and LP investors and AI fund managers. These relationships differ in the difficulty of specifying the principal's objectives, the difficulty of evaluating the agent's actions and outcomes, the losses caused by errors, and the possibilities for intervention. The paper examines how these differences change the governance challenges even though all these relationships involve delegation to AI.

The paper makes three contributions. First, by comparison with theories concerned with human agents, it clarifies the baseline assumptions and sources of the principal–AI agent problem. Second, it distinguishes behavioral divergence, governance costs, and residual loss, and presents an analytical framework connecting instructions, monitoring, the allocation of authority, and intervention. Third, it applies this framework to corporate governance involving AI managers and further identifies problems common to different delegation relationships and challenges specific to each. In doing so, it provides a basis for treating individual problems of delegation to AI as comparable theoretical questions.

The remainder of the paper begins with the foundations, reviewing the conventional principal–agent problem and agency theory. It then defines the principal–AI agent problem and presents the components of AI agency theory. Next, it examines possible forms of corporate governance in the relationship between shareholders and AI managers and considers problems that may arise in the other four delegation relationships. Finally, it summarizes the theoretical implications, scope of application, and directions for future research.

2 The Principal–Agent Problem and Agency Theory

2.1 Delegation Relationships and the Principal–Agent Problem

Principal–agent theory analyzes relationships in which one actor entrusts another with acting on its behalf. Ross (1973) understood an agency relationship as one in which one party acts for another in a particular domain of decision-making. The actor who delegates tasks or decisions is the principal, and the actor who undertakes them is the agent. Shareholders and managers, and employers and workers, are representative examples of this relationship.

Delegation provides a means of using expertise or information that the principal does not possess and obtaining the benefits of a division of labor. At the same time, granting the agent room for judgment raises the question of whether the agent's behavior will coincide with what the principal desires. It is therefore necessary to distinguish the existence of an agency relationship from the emergence of an agency problem. The former concerns who delegates what to whom; the latter concerns the difficulty of realizing behavior desirable to the principal within that relationship.

The principal and agent are also roles in the delegation relationship being analyzed, rather than fixed attributes of individuals. For example, a manager is an agent in relation to shareholders but becomes a principal when delegating tasks to workers. In each relationship considered here, this paper likewise makes clear whose objectives serve as the reference point and which tasks and judgments are delegated.

2.2 Conflicts of Interest, Information Asymmetry, and Risk

At the center of conventional agency problems are conflicts of interest between the principal and the agent and information asymmetry that prevents the principal from fully understanding the agent's behavior or ability. Agents act with regard not only to the principal's outcomes but also to their own compensation, the burden of effort, and discretion. Pursuing self-interest does not necessarily imply misconduct, but it creates the possibility that choices desirable to the agent differ from those desirable to the principal.

Moral hazard and adverse selection are distinguished as problems associated with information asymmetry. Moral hazard concerns choices of actions or effort that the principal cannot sufficiently observe. Adverse selection concerns the inability to adequately ascertain private information, such as ability, when contracting with or selecting an agent. For example, checking whether sufficient effort is supplied after hiring and determining whether the ability claimed at hiring is actually present require different governance responses (Eisenhardt, 1989).

In standard contract design, the principal chooses terms in anticipation of how the agent will respond to the contract. Two constraints are important: the participation constraint, under which accepting the contract is no less favorable to the agent than an alternative option, and the incentive compatibility constraint, under which the behavior the principal wishes to induce is implemented as the agent's own choice (Ross, 1973;

Holmström, 1979). Merely writing the principal's desired behavior into a contract does not guarantee that it will occur.

Furthermore, outcomes are not determined solely by the agent's behavior. Uncertainty about demand, market conditions, and other factors affects outcomes, making it difficult to infer actions or effort perfectly from observed results. Performance-based compensation can encourage behavior, but it also imposes on a risk-averse agent the burden of fluctuations beyond the agent's control. Contract design must therefore consider the relationship between the provision of incentives and risk sharing (Holmström, 1979).

2.3 Agency Costs and Residual Loss

Jensen and Meckling (1976) defined agency costs as the sum of monitoring expenditures by the principal, bonding expenditures by the agent, and residual loss. This can be expressed simply as follows.

$$AC = M + B + R \tag{1}$$

Here, M denotes the costs borne by the principal to understand and control the agent's behavior, while B denotes the costs borne by the agent to guarantee that the agent will not take actions that harm the principal or that compensation will be provided if harm occurs. R is the monetary equivalent of the welfare loss suffered by the principal due to behavioral divergence that remains despite these measures.

This classification distinguishes residual loss from expenditures to reduce it. Incurring excessive costs in an attempt to eliminate all divergence is not necessarily desirable for the principal. Taking the delegated tasks and their benefits as given, expenditures on governance and the losses that remain must be evaluated together. If the scope of delegation or organizational forms are also being compared, the benefits obtained from each must be considered as well.

This paper also does not classify the entirety of an agent's salary or the purchase price of a service as agency costs. It distinguishes the costs required for ordinary task execution from the costs of addressing delegation problems and residual loss. When assessing residual loss, the benchmark actions or outcomes must also be made explicit; losses should not be attributed to agency problems merely because performance was poor.

2.4 Incentives, Monitoring, and the Division of Decision-Making

One major governance instrument for addressing agency problems is incentive design. Performance-based compensation and similar arrangements seek to make behavior desirable to the principal worth choosing for the agent as well. However, if desirable outcomes cannot be measured adequately, linking compensation strongly to performance can create other distortions.

Holmström and Milgrom (1991) showed that when an agent performs multiple tasks, compensation affects not only the total amount of effort but also its allocation across tasks. When effort on one task competes with effort on another and only easily measured outcomes are reflected in compensation, important but difficult-to-measure tasks may be neglected. Governance therefore involves not only the strength of compensation incentives but also the design of what to evaluate and which tasks to combine under a single agent.

Another instrument is monitoring. Information about actions and outcomes is acquired and used for evaluation and control. The concept of monitoring in Jensen and Meckling (1976) includes not only observation but also control through budget constraints and operating rules. Monitoring in conventional theory is therefore not limited to deterring behavior through an awareness of being watched.

Governance is also achieved by dividing the decision-making process. Fama and Jensen (1983) divided decision-making into initiation, ratification, implementation, and monitoring, distinguishing decision management, which encompasses initiation and implementation, from decision control, which encompasses ratification and monitoring. How to allocate roles between the actor carrying out tasks and the actor approving and evaluating those decisions is an organizational issue in managing agency problems. This perspective provides a basis for analyzing not only compensation for agents but also who decides what and who reviews those decisions.

2.5 Incomplete Contracts and the Allocation of Authority

In contracts supporting delegation, it is difficult to specify in advance who should do what in every possible future situation. The cost of describing necessary actions and the difficulty of verifying performance leave some matters outside the contract. The

economics of incomplete contracts analyzes how the allocation of authority over decisions and resources affects the parties' behavior in such circumstances.

Grossman and Hart (1986) understood residual rights of control as the authority to decide how assets are used in ways not explicitly specified in a contract. Through the allocation of these rights, asset ownership affects ex post bargaining and ex ante investment incentives. Concentrating authority in one party also reduces the other party's authority, so integration or stronger control does not always improve efficiency. Who can address contractual gaps is also part of the design of a delegation relationship.

Furthermore, possessing authority is not the same as actually being able to influence decisions. Aghion and Tirole (1997) distinguished formal authority, the right to decide, from real authority, effective influence over decisions. Even when a superior has the final decision right, a subordinate's proposal may be adopted unchanged if the subordinate holds the information needed for judgment and the superior cannot acquire sufficient information. The allocation of authority must therefore be considered together with arrangements for acquiring and communicating information.

The fact that drafting and enforcing contracts is itself costly also falls within the scope of conventional agency theory. Fama and Jensen (1983) discuss the costs of structuring, monitoring, and bonding contracts among parties with divergent interests, as well as losses that remain because the costs of complete enforcement exceed its benefits. This paper retains these existing arguments when analyzing delegation to AI.

2.6 Stewardship Theory and Its Connection to This Paper

Even for human agents, it is not always necessary to assume strong conflicts of interest with the principal. Stewardship theory, as developed by Davis, Schoorman, and Donaldson (1997), addresses relationships in which agents identify with organizational goals and find intrinsic value in achieving them. In such cases, trust and empowerment that support agents' contributions are important. Appropriate governance depends on the agent's motivation and organizational circumstances.

However, the alignment of motivation assumed in stewardship theory must be distinguished from this paper's assumption that AI does not pursue private interests of its own. Human stewards seek to act in line with the objectives of the principal or organization while possessing their own motivations and value judgments. In contrast, the AI

agent in this paper is specified as an actor that operates according to externally supplied objectives. This setting removes the problem of prioritizing private interests of its own, but does not guarantee that the supplied objectives adequately reflect the principal's intentions.

Thus, undesirable outcomes for the principal can arise even when AI does not pursue private interests of its own and faithfully follows the specified objectives. As Ante (2026) argues, a gap between the objective intended by the principal and the objective actually specified can cause delegation losses. The issue here is not the agent's pursuit of private interests contrary to the principal's objectives, but the inadequate representation of the principal's intentions as the objectives and constraints that guide the agent's behavior.

This distinction also bears on the choice of governance instruments. The absence of the pursuit of private interests is not, by itself, sufficient grounds for judging that less oversight and greater discretion are appropriate. The appropriate form of delegation depends on how far objectives and constraints can be specified, how well actions and outcomes can be evaluated, and whether correction or interruption is possible when problems arise. This paper analyzes the presence or absence of motivational conflict separately from these conditions for governing delegation.

On this basis, the next section examines delegation to AI without assuming the pursuit of private interests of its own, while retaining conventional agency theory's discussions of cost structures, monitoring, and the allocation of authority. By clarifying how divergence between the principal's intentions and the AI's behavior arises and how governance instruments act upon it, the paper presents the components of the principal–AI agent problem and AI agency theory.

3 The Principal–AI Agent Problem and AI Agency Theory

3.1 Defining AI Agents and Delegation Relationships

3.1.1 AI Agents

This paper defines an AI agent as an AI system with the autonomy to select and execute actions toward a given objective using information obtained from its environment. Auton-

omy here means that the system can select individual procedures and actions under given conditions without a human specifying each one in turn. This definition establishes the analytical scope of the paper by drawing on the autonomy, reactivity to the environment, and goal-directed proactivity identified by Wooldridge and Jennings (1995).

The environment includes not only the physical world but also digital environments such as documents, databases, and business systems. For example, a system that selects the information needed for a research objective, conducts searches, and decides what to do next on the basis of the results falls within this definition. Even if human approval is required before important actions are executed, autonomy exists when there is room to choose other procedures and actions. Fully unattended execution is therefore not a definitional requirement.

The object of analysis is not limited to a model itself, but includes a system operated by combining a model with instructions, necessary memory, external tools, execution authority, and other components. It need not possess all these components. Even with the same model, different combinations of information and authority can lead to different ranges of behavior and outcomes. The ability to choose means autonomously is also distinguished from possessing objectives or private interests of one's own.

3.1.2 The Principal–AI Agent Relationship

This paper defines a principal–AI agent relationship as a relationship in which a human or organization entrusts an AI agent with performing tasks to realize its own objectives and delegates a degree of discretion over judgment or execution in carrying them out. The actor entrusting the tasks is the principal, and the AI system performing them on the principal's behalf is the AI agent. This definition applies Ross's (1973) conception of agency, in which one party acts for another in a particular decision domain, to delegation to AI.

For such a relationship to exist, AI must be entrusted with some judgment or task execution on the principal's behalf, beyond merely being used. The delegated discretion need not cover the entire task. For example, when AI is entrusted with researching suppliers and selecting candidates while a human gives final approval for an order, a principal–AI agent relationship exists within the delegated scope.

The principal and AI agent are roles in an individual delegation relationship. Analysis requires making explicit whose objectives serve as the reference point, which tasks are entrusted, which judgments and actions are delegated to AI, and who oversees and intervenes. The existence of an agency relationship is distinguished from the emergence of a problem within it. This paper calls the difficulties of realizing behavior consistent with the principal's intentions in this delegation relationship the principal–AI agent problem.

3.1.3 Baseline Analytical Assumptions

In the baseline setting, AI is assumed to act on objectives formed and set externally and not to pursue independent private interests. This assumption is not a factual assertion about all existing AI systems. Nor does it mean that AI always follows the user's instructions or fully understands the principal's intentions. The user's explicit instructions may differ from the objective that actually guides behavior, including the effects of training and system-level settings. The latter is referred to below as the effective objective. This does not, however, assume that all AI systems strictly maximize a single objective function.

Ante (2026) distinguishes the absence of independent conflicts of interest, fidelity to the effective objective, alignment of that objective with the principal's intentions, and realized outcomes that fulfill those intentions. This paper retains this distinction. Even in the absence of private interests of its own, a gap can remain between the principal's intentions and the effective objective. Moreover, even an objective aligned with those intentions does not guarantee its implementation or the resulting outcomes.

3.2 AI Agency Slack and Its Sources

This paper calls divergence between the principal's intentions and the AI's behavior attributable to constraints on delegation—such as communicating objectives, evaluating behavior, setting authority, and intervening when necessary—AI agency slack. Divergence here does not encompass every difference between the procedures the principal imagined and those selected by the AI. If multiple actions satisfy the principal's objectives and constraints equally well, their differences need not be treated as losses. What matters is divergence that is significant in light of the principal's objectives and constraints.

To analyze this divergence, the paper distinguishes the intentions the principal wishes to realize, the instructions expressing them, the AI's effective objective, actual behavior, and realized outcomes. The principal may fail to reflect intentions adequately in instructions, or the instructions may not be adequately reflected in the effective objective. Furthermore, external conditions and chance affect outcomes, so bad outcomes may occur even when behavior accords with intentions. Poor outcomes alone therefore do not establish the existence of slack.

The sources of slack can be classified into the following three categories. These are diagnostic categories and may overlap in a single case.

Source	Description	Example when delegating sales expansion
Underspecification	Objectives, constraints, or responses to exceptions needed to realize intentions are absent from instructions	Pricing is delegated without communicating the principal's intended minimum profit margin or procedures for checking exceptions
Misspecification	Specified objectives or evaluation metrics do not adequately represent the principal's intentions	The principal wants sustained profits, but short-term sales alone are used as the criterion of success
Noncompliance with specifications	Explicit conditions are not reflected in actual behavior	Approval before discounting is required, but discounts are implemented without approval

The existence of unspecified elements does not by itself imply a loss. If those elements are irrelevant to the actual decision, or if AI can appropriately fill the gaps from context or available rules, behavior consistent with intentions may be realized. Conversely, even detailed instructions can produce undesirable behavior through faithful execution if the specified objective misrepresents the principal's intentions. What matters is not the length of instructions but the adequate communication of objectives and constraints important for judgment.

The causes of noncompliance with specifications must also be distinguished. Noncompliance here concerns conditions explicitly stated by the principal and can occur even when AI follows its effective objective. Failures due to limited computational or perceptual capabilities require different responses from failures due to deficiencies in instruction priorities, approval procedures, or authority management. This paper assesses the former separately as general performance problems and focuses on the latter insofar as they are

attributable to the design and operation of delegation. When capabilities are limited, a failure to set an appropriate execution scope or conditions for consulting a human is treated as a problem of designing delegation to address those limitations.

In principle, the objectives and constraints used to assess slack are established before outcomes are observed. If the principal's intentions are unclear, clarification or consensus formation comes first, and any remaining ambiguity is recorded. Comparisons make explicit the conditions concerning capability, information, and the external environment, without demanding perfect hindsight from AI. This procedure distinguishes divergence attributable to delegation from general performance limitations, environmental uncertainty, and subsequent changes in the principal's evaluation.

3.3 The Problem of Objective Specification and Its Relationship to Existing Theory

Delegation can fail even with AI that does not pursue private interests of its own if the principal cannot adequately specify the desired objective. Sections 4 and 5 of Ante (2026) emphasize the distinction between acting in accordance with an effective objective and that objective adequately representing the principal's intentions. Under these conditions, the central constraint is how far objectives can be specified and their alignment with intentions verified, rather than how to restrain the pursuit of private interests.

The problem of representing the principal's objective only imperfectly is not unique to AI, however. Baker (1992) showed that when the principal's objective cannot be used directly in a contract and another performance measure is used to guide the agent's behavior, differences in the effects of actions on the objective and on the measure create inefficiency. This problem persists even when the agent is risk neutral. This paper retains the distinction between objectives and measures and examines how divergence arises in delegation to AI without assuming private interests of its own.

As reviewed in Section 2, existing theories of contracts and organizations address the costs of drafting contracts, acquiring information, dividing decision-making, and allocating authority. This paper therefore does not claim that existing theory cannot address objective specification or control. AI agency theory is a framework that retains those insights while clarifying the assumptions and channels through which they apply to AI.

The objects of analysis of related concepts must also be distinguished. Objective misspecification describes the relationship between the principal's intentions and the specified objective. Behavior that exploits gaps in specifications to improve a performance measure is one manifestation of such misspecification or incompleteness. By contrast, AI agency slack describes behavioral divergence within a delegation relationship and is distinct from the resulting loss to the principal. Expenditures to reduce that loss are identified separately as well. In separating the properties of objectives, behavior, losses, and the costs of addressing them, this paper follows the classification of neighboring concepts in Section 3 of Ante (2026).

3.4 AI Agency Costs

Addressing AI agency slack is costly, and losses may remain afterward. This paper defines AI agency costs as the governance costs required to prevent, detect, and correct slack, together with the principal's expected loss that remains. Retaining Jensen and Meckling's (1976) distinction between expenditures and residual loss, it classifies these costs analytically as follows.

$$AC_{AI} = S + M + I + R. \tag{2}$$

Here, S is the specification cost of creating and updating objectives, constraints, exception handling, evaluation criteria, and related elements. M is the cost of observing information use, behavior, and outcomes and verifying them against the principal's intentions. I is the cost of resources directly devoted to control and intervention, such as setting authority, approval, correction, and interruption. R is expected loss, including both losses from slack that remains after governance and benefits forgone because governance measures cause delays or mistaken interruptions. A common benchmark is used to compare governance arrangements, while losses from ordinary operational uncertainty and general performance limitations are distinguished. This equation is the paper's own cost classification; it neither reproduces Ante's equation nor implies that each component is determined independently.

Specification cost in Ante (2026) includes the costs of designing and revising objectives, verification costs related to objective specification, and residual loss caused by the gap. In this paper's classification, these correspond to the portions of S, M, and R attributable

to objective specification. This paper additionally makes explicit governance conditions concerning the feasibility of approval, restrictions on authority, correction, and interruption. Because the two classifications overlap, specification cost is not added separately to AI agency costs.

Costs must be measured without counting the same activity or loss more than once. For example, time spent discovering a problem in review is classified as M, while time subsequently spent rewriting instructions is classified as S. Work and equipment directly required for intervention are included in I, while operational benefits forgone through waiting for approval or mistaken interruption are included once in R relative to the benchmark. Ordinary AI usage charges and general computing costs are likewise not classified in their entirety as agency costs.

Bonding expenditures by AI itself are not included as a separate component in the baseline setting. This is because AI without private interests of its own is not assigned the role of an actor purchasing guarantees to restrain its own opportunism. Costs may nevertheless arise when a provider offers a guarantee or a human supervisor commits to performance. These are analyzed separately as contractual relationships among the humans and organizations surrounding AI.

The purpose of governance is not to eliminate slack at any cost. Given the delegated tasks and their benefits, the sum of governance costs and residual loss must be compared. When choosing the scope of delegation or the system itself, ordinary execution costs and benefits from speed, scale, specialized processing, and other features must also be evaluated. The same governance measure may reduce losses while increasing the burden of checking and delays.

3.5 Specification, Monitoring, and Intervention

The first governance instrument is specifying and updating objectives and constraints. In addition to what should be achieved, it defines which means are permitted, what must not be sacrificed, and whom to consult when a judgment cannot be made. If it is difficult to exhaustively describe all situations in advance, procedures for consultation and updating when exceptions are discovered can be established. It is important to combine detailed instructions with arrangements for responding to unspecified situations.

Hadfield-Menell and Hadfield (2019) connect AI problems with the way incomplete human contracts are supported not only by explicit clauses but also by external structures such as law, norms, and conventions. This paper uses that perspective to consider arrangements linking organizational policies and operating rules to AI decisions. Rules must not merely be stored as documents; they must be consulted when needed and reflected in judgment and execution.

The second instrument is monitoring and verification. This involves understanding the information consulted by AI, the actions selected, the course of execution, and outcomes, and evaluating them against the principal's objectives and constraints. Because the success metric itself may be mistaken, checking only for improvements in that metric is insufficient. For example, when evaluating sales growth, other aspects important to the principal, such as profit margins and continuing customer relationships, also become subjects of verification.

In this baseline setting, the central channel of monitoring is not AI restraining its pursuit of private interests because it is aware of being monitored. Rather, it is detecting divergence through observation and using that information to change instructions, authority, or execution methods. The paper retains the broad conception of monitoring described in Section 2 while specifying which information helps which corrections in delegation to AI.

The third instrument is the allocation of authority and intervention. It specifies how far AI may decide and act on its own, which actions require prior approval, and who may correct or stop execution. Given the distinction between formal and real authority in Aghion and Tirole (1997), merely retaining human authority to approve or stop actions is insufficient. The ability to understand the information needed for judgment and the time to intervene are also necessary. Applied to AI, this perspective focuses analysis on whether approval entails substantive evaluation and whether execution can be changed after a problem is identified.

These instruments operate in combination. Even with verification information, it is difficult to correct ongoing behavior if intervention is impossible. That information can, however, be used to improve subsequent delegation or instructions. Restricting the scope of execution in advance can also sometimes reduce the need for subsequent monitoring.

Specification, verification, and intervention are therefore not always complementary; it is necessary to examine where combinations are required and where substitution is possible.

3.6 The Scope of Delegation, Allocation of Computing Resources, and Interruptibility

In this paper's baseline setting, AI agents are not assigned independent private utility or preferences over compensation. Incentives that reward AI itself and link its private payoff to the principal's outcomes therefore do not exist in this analysis. Reward signals during training are also treated as technical mechanisms for shaping behavior and distinguished from incentive contracts between a principal and AI. Under this assumption, what the principal designs are the conditions of delegation: how far judgment and execution are delegated, how much computing capacity is allocated, and under what conditions execution can be interrupted.

First, control over the scope of delegation is exercised by defining the tasks entrusted to AI and the permissible range of judgments and actions within those tasks. For example, researching suppliers, selecting candidates, placing orders, and making payments can each be delegated separately. Broad discretion can be granted for research and candidate selection, while orders are subject to monetary limits and payments require the principal's approval. Alonso and Matouschek (2008) analyze delegation when contingent monetary transfers are unavailable as a problem in which the principal chooses a set of permissible decisions. This paper retains that perspective of designing the available choices of action. It does not, however, transfer that study's assumptions about agent preferences or incentive compatibility directly to AI.

Narrowing the permissible range can reduce the scope for undesirable actions, but may also prevent appropriate responses to circumstances. Expanding the range of actions requiring approval increases the principal's checking burden and delays execution. The scope of delegation must therefore be determined in light of the magnitude of losses from errors, the evaluability of judgments, the reversibility of actions, and the urgency of response. Strong past performance likewise does not automatically justify entrusting AI with different tasks or greater authority. When the scope is expanded, capability and governance conditions for the new tasks must be assessed anew.

Second, the allocation of computing resources concerns the resources devoted to AI task execution and verification. Computing resources are not rewards for AI, but inputs used for information searches, reasoning, comparisons of alternatives, and checks of results. Even under the same execution authority, the use of computing resources can change the quality of judgment and processing time. Conversely, increasing computing resources does not by itself expand the range of tasks or external operations AI is allowed to perform. Resource allocation affecting capability and authority allocation affecting permissible actions must be designed separately.

Additional computation may reduce errors, but its effect depends on the objective pursued and the processes to which resources are devoted. If the effective objective diverges from the principal's intentions, more computation may enable more intensive pursuit of that objective. Conversely, excessive resource constraints may prevent necessary searches or verification and reduce the quality of judgment. The principal determines allocations by comparing the expected benefits of additional computation with computing costs and effects on response time. In doing so, computing costs for ordinary task execution are distinguished from additional verification costs for detecting and correcting slack, and the former are not included in their entirety in AI agency costs.

Third, interruptibility means that the principal can stop AI processing or external operations during execution, assess the situation, and choose among correction and resumption, reduced authority, handover to another actor, or termination of delegation. Instructions and authority considered appropriate at the outset may become inappropriate as new information arrives or the environment changes. Interruptibility is understood as a mechanism for revising the conditions under which delegation continues in response to such changes.

Formally possessing the authority to interrupt is different from actually being able to interrupt when necessary. Given the distinction between forms of authority in Aghion and Tirole (1997), discussed in Section 3.5, interruption requires not only the information and time needed for the decision but also a means of implementing that decision. In this paper's application to AI, a governance condition is that the principal can suspend processing or authority over external operations, rather than relying solely on instructions to the AI to interrupt. However, even if execution can be stopped, payments already

made or information already disclosed may not be reversible. Interruptibility must be distinguished from the reversibility of the consequences of actions.

Interruption is not always desirable. In addition to losses from overlooking an anomaly and continuing, there may be losses from mistakenly stopping an unproblematic process, costs of checking and recovery, and the burden of handing work over to a human or another actor. Decisions to interrupt therefore require a comparison, using information available at the time, between the expected net benefits of continuation and those of interruption followed by correction and resumption, handover, or other responses. Under the cost classification in Section 3.4, resources directly devoted to stopping or handing over are counted in I, while benefits forgone through delay or mistaken interruption are counted in R, each only once.

These three elements are distinct but interrelated choices in designing delegation. When resources for verification are scarce and interruption cannot occur in time, the need to restrict execution in advance may increase. Conversely, broader discretion may be feasible when judgments can be adequately verified and intervention remains possible if problems arise. This paper analyzes these elements not as ways to change AI's motivation, but as ways to define the actions available, the resources available for execution and verification, and the possibilities for reconsidering delegation. Compensation and incentives for the humans and organizations providing and overseeing AI are treated as separate contractual relationships from delegation to AI itself.

3.7 Analytical Implications and Scope of Application

The preceding analysis yields several testable questions about delegation to AI. The following are conditional hypotheses informed by the predictions in Section 5 of Ante (2026), rather than propositions whose general validity has been proved in this section.

First, when important conditions related to the principal's intentions are unspecified and those conditions actually affect decisions, improving specification is expected to reduce residual loss. If the required work costs more than the losses reduced, however, total AI agency costs increase. The effect of more detailed instructions is evaluated through both improvements in behavior and the costs of drafting and updating them.

Second, when previously unproblematic unspecified conditions become important in a new environment and are not appropriately filled in, slack and residual loss are expected

to increase. Ante presents this as a prediction concerning environmental novelty. This paper distinguishes this effect from failures due to reduced predictive ability or perceptual accuracy. Continuing to use the same model does not by itself hold capability constant; comparisons must also consider available information, task difficulty, and changes in governance.

Third, multiple AI systems using a common model, instructions, or evaluation criteria may make similar errors in common situations. If these commonalities affect failures beyond differences in individual tasks and environments, increasing the number of AI systems will not sufficiently diversify the risk of loss. The use of a common model alone, however, does not establish correlated failures. Correlation alone also does not change the additivity of individual expected losses. For total losses to exceed those under independence, an additional mechanism is needed through which simultaneous errors amplify damage, for example through price changes or liquidity shortages. Ante (2026) likewise distinguishes correlation in failures from superadditivity of losses.

Fourth, the desirable combination of governance instruments depends not only on the specifiability of objectives but also on the evaluability of behavior, the time available for intervention, the reversibility of errors, and the benefits of delegation. Approval may be effective for errors that can be detected through verification and corrected before execution. For tasks whose main benefits are lost while awaiting approval, however, comparison with other arrangements, such as prior authority limits or interruption in exceptional cases, is necessary. This paper seeks to explain the conditions under which oversight and autonomy should be combined, rather than uniformly strengthening either.

The framework has two boundaries. The first concerns AI with independent private objectives. Ante (2026) identifies persistence of behavior after the specified objective is changed and the absence of instrumental benefit of that behavior for the specified objective as criteria for considering independent objectives. Persistence alone, however, cannot distinguish these from higher-level instructions, learned tendencies, or incomplete implementation of a change. If evidence warrants assuming independent objectives, the baseline setting must be extended to address both objective specification and conflicts of interest.

The second boundary concerns who determines the principal's intentions themselves. Within organizations, shareholders, managers, clients, and others need not share the

same objectives, and what AI should be instructed to do may not be settled in advance. This problem of consensus formation precedes the problem of communicating agreed objectives to AI. Furthermore, even when the principal's objectives are achieved, third parties outside the scope of delegation may suffer losses. Such external losses must be evaluated separately, rather than obscured by including them in the principal's residual loss.

The next section applies this framework to the relationship in which shareholders are the principal and AI entrusted with managerial decisions and business operations is the agent. It examines how shareholders' objectives are reflected in managerial instructions, who evaluates AI decisions, which actions are approved, and who modifies or stops execution when necessary. In this way, corporate governance involving AI managers is considered as a set of arrangements for objective formation, specification, oversight, authority allocation, and intervention.

4 AI Agents and Corporate Governance

This section applies the framework in Section 3 to a relationship in which shareholders delegate managerial decisions and business operations to AI. It first reviews what produces losses from delegation in the conventional shareholder–manager problem and which governance instruments have been considered. The implications of delegating management to AI become clear by distinguishing which mechanisms change and which problems remain in analyses premised on human managers.

The separation of shareholders and managers does not itself imply inefficiency. Those with the knowledge and judgment needed for management are not necessarily the same actors as those who supply funds and bear risk. Fama and Jensen (1983) argue that while a division of labor between decision-making and residual risk bearing has benefits, arrangements for decision control are needed when decision-makers do not personally bear most of the consequences. The object of analysis is therefore not the elimination of separation itself but the design of organizations that obtain its benefits while managing the problems associated with delegation.

Jensen and Meckling (1976) analyze how outside shareholders affect managerial behavior when managers obtain private benefits in addition to monetary returns. As managers'

ownership shares decline, they bear a smaller fraction of the reduction in firm value caused by using corporate resources for private benefits. Moreover, if managers personally bear the burden of effort while some of the returns accrue to outside shareholders, effort devoted to seeking profitable business opportunities may also be insufficient. This argument explains behavioral divergence through the relationship between private benefits and costs and the manager's stake in firm value, rather than through the manager's character.

The paper's basic analysis of outside equity, however, assumes a single owner-manager and, among other conditions, that outside equity carries no voting rights. Managerial ownership alone therefore cannot determine the quality of governance in every firm. What should be retained analytically is the question of who receives the benefits and bears the costs of decisions and how that distribution changes behavior.

Conflicts of interest also appear in the allocation of funds. Jensen (1986) argues that when free cash flow exceeds the amount required for favorable investment opportunities, a conflict may arise between returning funds to shareholders and maintaining or expanding the resources controlled by managers. If managers benefit from organizational size or discretion, they may choose low-return investments over distributions to shareholders. The problem here is not investment failure in general, but distortions in capital allocation attributable to managers' private benefits.

These theories do not justify treating a deterioration in corporate performance immediately as an agency problem. Losses caused by changes in demand or technological uncertainty are distinguished from behavioral divergence caused by managerial self-interest. This distinction must be maintained when comparing delegation to AI.

Share ownership and performance-based compensation have been examined as means of bringing shareholders' and managers' interests closer together. Their design, however, is not simply a matter of strengthening the link between pay and performance. Holmström (1979) analyzes contracts when actions cannot be directly observed and demonstrates the relationship between incentives and risk sharing. Outcomes include fluctuations beyond managerial control, making it costly to impose performance risk on risk-averse managers. Observations that provide information about actions in addition to existing performance measures can improve contracts under certain conditions, but when acquiring or using such information is costly, the net benefits must be assessed separately.

Managers' responsibility for multiple tasks also complicates compensation design. Holmström and Milgrom (1991) show that compensation changes not only total effort but also its allocation among tasks. When easily measured and difficult-to-measure tasks compete for effort, strong incentives for the former may harm the latter. In this analysis, the division of jobs and restrictions on behavior are governance choices alongside compensation. It is therefore inappropriate to characterize conventional theory as a theory solely of monetary rewards.

Empirical research also indicates that the existence of performance-based compensation alone is insufficient to assess a contract's appropriateness. Bertrand and Mullainathan (2001) report that executive compensation in U.S. firms responds to performance fluctuations attributable to factors considered outside managerial control, such as oil prices, and that this response is associated with the strength of governance. Governance structures themselves are not randomly assigned, however, so the observed association cannot simply be interpreted as a causal effect of governance. Moreover, once multitask incentives are considered, compensation responses to exogenous fluctuations are not necessarily always inefficient (Holmström, 2017). Measuring outcomes and evaluating contracts that induce desirable behavior must be undertaken separately.

A governance instrument alongside compensation is the division of the decision-making process. Fama and Jensen (1983) distinguish decision management, encompassing initiation and implementation, from decision control, encompassing ratification and monitoring. In firms with many shareholders, it is costly for every shareholder to check each managerial decision, so the control function itself is delegated. The board of directors is understood as an organizational body responsible for this control.

The existence of a board, however, is not the same as effective oversight. Adams, Hermalin, and Weisbach (2010) review boards' monitoring and advisory functions while emphasizing that their composition and behavior are determined endogenously in response to firm circumstances and relationships with managers. For example, higher performance among firms with more independent boards does not by itself establish the causal effect of increasing independence. Monitoring and advice also differ in their information requirements and relationships with managers, making it difficult to judge a board's overall effectiveness from a single characteristic.

Concentrated ownership also has benefits and costs. As reviewed by Shleifer and Vishny (1997), large shareholders capture a substantial share of improvements in firm value from monitoring and are therefore more willing than small shareholders to bear monitoring costs. Strong control, on the other hand, may create conflicts with minority shareholders. The interests of those monitoring managers need not coincide with the interests of all shareholders. Analysis must therefore examine not only the existence of a monitor but also that monitor's objectives and authority.

Fama (1980) analyzes discipline through the managerial labor market. If current behavior and performance change assessments of ability and thereby affect future compensation, incentives arise that do not depend solely on current performance-based pay. This mechanism depends, however, on conditions such as the market's ability to evaluate managerial contributions and the manager's concern for future assessments. The argument analyzes the division of management and risk bearing together with internal and external labor markets and monitoring arrangements.

Debt and corporate takeovers have also been analyzed as mechanisms for disciplining management. In Jensen's (1986) account, a commitment to future payments through debt may reduce the funds managers can freely allocate and restrain the waste of free cash flow. Debt also entails costs such as bankruptcy, however, and the same control need not suit firms with many favorable investment opportunities. Takeovers can provide an opportunity to change inefficient management, but can themselves also become a vehicle for managerial overinvestment.

These studies analyze the shareholder–manager problem through combinations of compensation, information, authority, ownership structure, and capital allocation. Building on this body of work, the present section examines the channels through which each instrument operates when management is delegated to AI.

This paper calls an AI agent entrusted with a defined range of managerial decisions and business operations to realize shareholders' objectives an AI manager. This is an analytical designation focused on management functions; it does not presuppose that AI has legal eligibility to serve as a corporate officer or can bear legal responsibility. Delegating only some functions, such as investment, pricing, cash management, or procurement, also falls within this section's scope for the tasks delegated.

As in Section 3, AI managers are not assigned independent private utility or preferences over compensation. The baseline setting therefore excludes the channel assumed for human managers whereby corporate resources are used to obtain private wealth or status. This paper likewise does not assume incentives that give AI shares or bonuses to align its interests with those of shareholders. Additional computing resources and an expanded scope of delegation are changes in the conditions of task execution, rather than rewards for AI.

Nevertheless, the separation between those making managerial decisions and those bearing the resulting economic losses remains. For example, if an objective gives excessive weight to sales growth, AI may expand unprofitable businesses without seeking private benefits. Although the outward appearance of overinvestment is shared with the human-manager case, the cause is objective misspecification rather than managerial self-interest. Introducing equity compensation cannot be said to eliminate this cause.

The relationship between conventional governance instruments and their application to AI can be summarized as follows.

The right-hand column of the table presents this paper's application of the assumptions in Section 3, rather than results that previous studies established for AI. For example, the discipline through future compensation discussed by Fama (1980) does not operate unchanged for AI that is not assumed to have preferences over such compensation. Similarly, when restrictions on AI authority reduce benefits, their channels of operation must be identified, such as losses from excluding useful options or waiting for approval.

Delegating management to AI first requires determining what counts as shareholders' objectives. An instruction to "act for shareholders" alone does not establish criteria for specific managerial decisions. Even when increasing firm value is adopted as the basic analytical objective, short-term accounting profits, sales revenue, and observed share prices cannot simply be treated as identical objectives.

Furthermore, when shareholders also value the effects of corporate activities beyond financial returns, the distinction between market value and shareholder welfare becomes relevant. Hart and Zingales (2017, 2022) argue that the two may diverge, for example when shareholders care about social effects and cannot readily separate corporate activities from those effects through their individual actions. If shareholders differ in preferences or judgments, procedures for reflecting those differences in corporate deci-

Table 1: Comparison of Governance Instruments for Human and AI Managers

Instrument	Operation for human managers	Application to AI managers (this paper's analysis)
Share ownership and performance-based pay	Link managers' returns to shareholders' interests. Entail risk bearing and distortions in performance measurement.	Not treated as incentives for AI itself, since preferences over compensation are not assumed. Objective and evaluation-criterion design are examined.
Boards and monitoring	Provide approval, oversight, and advice on decisions. Effectiveness depends on the allocation of information and authority.	Verify instructions, judgments, and actions, linking findings to objective updates, approval, correction, and interruption.
Monitoring by large shareholders	Facilitate internalization of monitoring benefits, but may create conflicts with minority shareholders.	Examine the objectives and authority of humans and organizations providing oversight. Agency problems among humans remain.
Managerial labor markets	Assessments of ability and contribution affect current behavior through future positions and compensation.	Discipline through future compensation is not assumed. Performance and judgment evaluations inform continued use or replacement.
Financial constraints through debt	Reduce freely allocable funds and may restrain overinvestment driven by private interests. Also entail bankruptcy and other costs.	Examined as constraints on spending authority and capital allocation, separately from deterring self-interest. Costs of preventing useful investments are also assessed.
Dismissal, takeovers, and replacement	Change management and affect behavior through the possibility of losing position or pay. Takeovers may themselves be distorted.	Interruption and replacement change subsequent processing and decisions. Handover and recovery costs are assessed without assuming a response to private losses.

Note: The human-manager column summarizes the studies cited in the text. The AI-manager column presents this paper's analysis based on the baseline assumptions in Section 3.

sions are also necessary. This argument illustrates the problem of forming objectives before instructions are given to AI.

This section does not assume that AI itself substitutes for this process of objective formation. Instead, objectives and constraints adopted through prescribed governance procedures serve as the criteria for evaluating delegation. It then clarifies who translates objectives into concrete terms, on what information they are updated, and who approves the changes. Conflicts among shareholders in forming objectives and incompleteness in communicating the resulting objectives to AI are treated as separate problems.

For example, when capital investment is delegated to AI, decisions must reflect not only profitability but also cash flow, the investment horizon, tolerance for loss, and relationships with existing businesses. If short-term profits alone are the evaluation criterion, useful research and development may be cut. Conversely, evaluating only business scale may lead to expansion without profitability. These are examples of objective misspecification considered in this paper, rather than predictions that AI will necessarily behave in these ways.

Because it is difficult to specify every future situation in advance, specifications include not only current decision criteria but also procedures for consultation and updating if assumptions change. For example, if demand prospects or procurement conditions change substantially, conditions for requesting a review can be set, rather than simply executing the initial investment plan. Governance of AI managers concerns not only the content of objectives but also the authority to change them.

Even when shareholders are the ultimate principal, they need not directly undertake all day-to-day instruction and verification. This section considers a structure in which shareholders entrust governance to a board or similar body that manages the AI manager's objectives, authority, evaluation, and continued use. This structure distinguishes delegation from shareholders to the governing body from delegation from that body to AI. The assumption that AI has no private interests of its own does not eliminate the human agency problem in the former relationship. If the humans drafting instructions choose evaluation criteria favorable to themselves, AI may fail to serve shareholders' objectives even when it faithfully executes those instructions.

Using Fama and Jensen's (1983) division of the decision process, delegation to AI can be designed separately for initiation, ratification, implementation, and monitoring. For

example, AI can be entrusted with searching for investment candidates and analyzing profitability, while the board retains approval of projects above a certain scale and AI executes approved orders within prescribed limits. Delegating management does not mean transferring all stages together to the same actor.

The scope of approval is defined in terms not only of monetary amounts but also of the magnitude of losses, the evaluability of judgments, the reversibility of actions, and the time available for response. Even small individual expenditures may accumulate into large outflows, making total spending over a period a subject of control alongside per-transaction limits. If managerial decisions are redelegated to subordinate AI systems, the scope of the original constraints must also be made explicit so that they are not lost through redelegation.

Oversight examines not only final outcomes such as profits or share prices but also the information and assumptions used at the time of decision, the actions selected, and approval histories. Good outcomes may reflect inappropriate actions rewarded by chance, while bad outcomes may follow appropriate judgments given the information available at the time. Evaluation must use criteria established in advance and information that was available at that point.

Explanations generated by AI are not the sole basis for verification. It is important for monitors to have access to information against which explanations can be checked, such as transaction records and original materials. If another AI system is used for verification, the possibility that it shares errors in information or evaluation criteria must also be examined. This applies the problem of failures arising from common mechanisms, discussed in Section 3, to monitoring within firms.

Furthermore, even with authority to approve or stop actions, effective control weakens when the time or expertise needed to understand information is lacking. Drawing on Aghion and Tirole's (1997) distinction between formal and real authority, this paper evaluates board authority together with access to information and the feasibility of intervention. The issue is whether important judgments can be evaluated and necessary changes made in time, rather than merely increasing the number of approvals.

Allocating computing resources to an AI manager means allocating inputs that support managerial decisions. Searching for candidate projects, comparing scenarios, checking calculations, and verifying execution results each require resources. Distinguishing resources

for ordinary processing from those for additional verification of important judgments removes the need to devote the same amount of computation to every task. Resources are allocated by comparing the prospect that additional computation will improve decisions with its cost and time requirements.

More computation does not, however, guarantee an appropriate objective. If the objective is mistaken, additional computation may advance search or optimization toward that objective. Conversely, overly restrictive computing resources may prevent useful comparisons or verification. Computing budgets are therefore determined in relation to objectives, processing tasks, and necessary verification, rather than treated as rewards or sanctions for AI. Computing budgets and authority to spend corporate funds are also managed separately. The ability to use substantial resources for analysis does not confer permission to execute large investments unilaterally.

Interruptibility is a mechanism for reconsidering the conditions for continuing delegation to AI during execution. A board or operational manager must be able to stop processing or external operations when problems are suspected and choose among revised instructions, reduced authority, handover to a human, or replacement with another system. As in Section 3, a governance condition is that the delegating party can suspend authority or execution pathways without relying solely on AI's compliance with instructions.

Dismissal of a human manager may change behavior not only through the management changes brought by actual replacement but also through the possibility of losing future position or compensation. No such response to private losses is assumed for the AI manager in this paper. Interruption and replacement operate by changing subsequent processing and decisions. Moreover, interruption does not necessarily cancel existing contracts or completed transactions, so the ability to stop must be distinguished from the reversibility of outcomes.

Interruption entails costs from operational delays, assessing the situation, recovery, and handover. If AI is deeply integrated with internal information and business procedures, replacing it with another system may not be easy. Preparing for business continuity and the transfer of records after interruption is therefore an investment in making formal replacement authority a practically available option. Such preparation is also costly,

however, so it should be assessed according to the importance of delegated tasks rather than required at the same level for all tasks.

When AI is supplied by an outside firm, another contractual relationship is added. Providers may have their own interests in fees, usage volumes, and contract renewal. Divergent interests between sellers and buyers alone do not establish an agency problem, however. An agency problem arises, for example, when a client firm delegates resource allocation or system operation to a provider and cannot fully observe those choices, while a provider earning usage-based revenues can choose processing that is not commensurate with the client's benefits. Effective control by the client also depends on who can determine changes to models and settings, access to records, and conditions for stopping or migrating the system.

These are not incentive problems of AI itself but agency problems among humans and organizations: shareholders, boards, operators, and providers. Conventional governance instruments such as compensation contracts, information disclosure, monitoring, and contract renewal remain subjects of analysis in these relationships. The introduction of an AI manager cannot be understood as eliminating conflicts of interest throughout the firm.

The preceding discussion does not establish that replacing human managers with AI necessarily lowers a firm's agency costs. What the baseline setting excludes is behavioral divergence attributable to AI's own private interests. Objective specification, verification, authority management, interruption, and replacement nevertheless entail costs, and losses may remain after these responses. Agency problems involving the humans governing AI or the firms supplying it must also be evaluated.

Costs of delegation to AI are classified according to the specification cost S, monitoring and verification cost M, control and intervention cost I, and residual expected loss R in Section 3. Drafting and updating objectives and constraints correspond to S, verifying judgments and actions to M, and resources directly devoted to approval, interruption, and handover to I. Losses from divergence remaining after governance, together with benefits forgone through delays or mistaken interruptions, are included in R. No activity or loss is counted in multiple components, and computing costs required for ordinary task execution are distinguished as well.

Comparisons among human management, AI management, and combinations of the two consider not only governance costs but also the quality of judgment, processing speed, available expertise, and ordinary execution costs. The comparison concerns which combination delivers the highest expected net benefit for the same tasks and evaluation criteria. If business activities or organizational structures change at the same time as managers are replaced, not all changes in performance should be attributed to AI.

The framework in this section has the following conditional implications. First, the benefits of granting AI broad discretion may be greater in tasks where important objectives and constraints can be specified and behavior can be verified at relatively low cost. Second, prior approval or limits on the scope of delegation may be more important for tasks with highly ambiguous objectives or losses that are difficult to recover after execution. Third, when stopping or migration costs are high, the value of governance depends not only on performance at adoption but also on the conditions for reconsidering delegation. These are analytical hypotheses of this paper, rather than empirical findings establishing the superiority of AI management.

The central question inherited from conventional research on shareholders and managers is how to design an organization that obtains the benefits of entrusting management to others while aligning their decisions with shareholders' objectives. For AI managers, the baseline setting excludes the mechanism of restraining private interests of their own and instead analyzes objective formation and specification, the division of decision-making, resource allocation, and the feasibility of interruption and replacement. By also addressing the interests of the humans and organizations responsible for these functions, the shareholder–AI manager relationship can be situated within governance of the firm as a whole.

The next section considers delegation to AI in roles other than management and examines how the framework applies when the principal's objectives, delegated discretion, information available for evaluation, and conditions for intervention differ.

5 The Scope of AI Agency Theory and Comparison of Delegation Relationships

5.1 Differences in Roles and a Common Analytical Perspective

The preceding section examined the relationship in which shareholders entrust managerial decisions and business operations to AI. This section applies the framework in Section 3 to AI in the roles of worker, bureaucrat, professional, and fund manager. It clarifies how conventional agency problems change and which problems remain when AI performs these tasks in place of humans.

These names describe the functions entrusted to AI; they do not confer legal worker status, civil-service status, professional qualifications, or similar standing on AI. The four roles can also overlap. Examples include AI making professional judgments within a firm or a public organization entrusting asset management to AI. The object of analysis is therefore not determined by the AI's designation. Attention must be paid to individual delegation relationships: who entrusts which judgments and actions, and for what purpose.

The following analysis maintains the assumption in Section 3 that AI itself has neither independent private utility nor preferences over compensation. The effects on human behavior of expectations of pay or promotion, concern for reputation, or a sense of professional mission therefore cannot be used unchanged to explain AI behavior. Objective specification, evaluation of behavior, allocation of authority, and intervention when errors occur nevertheless remain issues in delegation to AI. Each subsection reviews the mechanisms identified by previous research and then applies this paper's framework to the relevant domain.

5.2 Organizations and AI Workers: Changing Tasks and Allocating Work

AI acting as a worker is an AI agent entrusted with judgments and tasks within a defined scope under organizational work objectives. In this relationship, it is important not only to specify tasks initially but also to change their content and priorities as circumstances evolve.

Simon (1951) formalized employment as a relationship in which, in return for a wage, a worker allows an employer to choose tasks within a range acceptable to the worker. Compared with a contract fixing the work in advance, the advantage is that specific tasks can be selected after circumstances become clear. This "area of acceptance," however, is determined by the worker's preferences and the wage offered. In this paper, which assumes no independent private utility for AI, it cannot be equated with the scope of authority granted to AI. What is retained is the perspective of distinguishing conditions fixed in advance, matters the organization can change later, and matters left to the judgment of the actor performing the work.

The allocation of effort across multiple tasks is also an issue. Holmström and Milgrom (1991) show that strongly linking compensation to performance on an easily measured task may concentrate effort on that task and reduce effort on competing tasks. Their analysis covers the division of jobs and restrictions on activities as well as compensation. With AI, too, using only the number of cases processed as an evaluation criterion may lead to neglect of accuracy or exceptions. This is not, however, because AI allocates effort in pursuit of higher wages. It arises because the objectives and evaluation criteria given to AI do not adequately represent the objectives of the work as a whole.

Accordingly, delegating work to AI requires defining not only the division of tasks but also who can change instructions, the priority among conflicting instructions, and the conditions for handing work over to others. For example, even if answering inquiries, proposing refunds, and issuing refunds are treated as a single customer-service workflow, each stage requires different information and authority. If the criterion for considering a case resolved is inappropriate, an increase in the number of cases processed by AI may merely shift the burden of corrections and complaints to another department. Evaluation must therefore consider effects on the whole workflow, not only the part directly handled by AI.

The scope of AI's role must also be checked when reading empirical research on AI assistance. Brynjolfsson, Li, and Raymond (2025) analyze the staggered introduction of an AI assistance tool among 5,172 customer-support agents at one firm. They report that access to the tool increased resolved cases per hour by about 15% on average. In this arrangement, however, humans could edit or decline the AI's suggestions. The study does not establish the effects of delegating the entire task to AI. Its main analysis also

uses a difference-in-differences approach, and interpreting the results causally requires assumptions such as parallel trends.

When evaluating the division of work between humans and AI, productivity improvements must be considered together with the human time and effort required to update instructions, verify work, handle exceptions, and arrange handovers. Insufficient AI skills and delegating tasks without accounting for those limitations are separate problems. The latter may be improved by revising the scope of delegation or establishing checking procedures. Such governance arrangements cannot, however, compensate for every deficiency in capability.

5.3 The Public Sector and AI Bureaucrats: Determining Objectives and Performing Tasks

AI acting as a bureaucrat is an AI agent entrusted with judgments about matters such as eligibility assessment, resource allocation, and working procedures under the objectives and constraints of an administrative organization. This cannot be understood simply as a relationship in which one politician instructs AI. There is a chain of delegation from citizens to political representatives, administrative organizations, and operational staff; multiple actors with different interests may also influence the same administrative activities.

Dixit (2002) analyzes public-sector organizations with attention to multiple objectives and principals, hierarchical delegation, and difficulties in measuring outcomes. A hierarchical relationship in which an actor is an agent to a superior and a principal to a subordinate must be distinguished from a relationship in which multiple principals influence a single agent. The interests of those receiving public services, those paying for them, and those concerned with their implementation need not coincide. The public sector must address both the determination of objectives and the execution of tasks in accordance with them.

Research on bureaucracy has also analyzed motivations other than monetary compensation. Prendergast (2007) shows that bureaucrats' own preferences over policy outcomes affect their effort to acquire information. Which preferences are desirable in a bureaucrat also depends on the relationship of interests with service users and the opportunities to correct errors. This mechanism assumes that humans derive private utility from policy

outcomes. It therefore cannot be interpreted here as implying that giving AI a sense of mission or empathy elicits effort. When AI judgments are biased, it is necessary to examine whether the source lies in specifications, data, evaluation methods, or authority.

Under this framework, delegating administrative work to AI first requires separating authority to approve or change objectives and constraints from authority to process individual cases. In assessing benefit applications, for example, the issues extend beyond processing volume and time to how mistaken denial and mistaken approval are evaluated, how supporting evidence is checked, how exceptions are treated, and what procedures allow requests for correction. If these are determined solely by internal evaluation criteria of the responsible department or provider, AI may depart from organizationally approved objectives even while following instructions. Improved computing capabilities do not resolve who should determine such criteria.

Those receiving public services are not always principals able to instruct AI directly. Even if an affected person notices a processing error, that information is difficult to use in oversight without a mechanism for reporting it and requesting review or correction. Examining only contested cases may also overlook problems affecting people who have not noticed an error or face high costs of lodging an objection. How to combine reviews initiated by objections with separate audits is therefore an issue.

Requiring prior human approval for every operation, on the other hand, may increase delays and checking burdens and reduce the benefits of public services. Automatic execution, ex post auditing, individual approval, and transfer to other staff must be used according to the clarity of decision criteria, the verifiability of judgments, the severity of harm from errors, and the ease of correction. Lower administrative costs are not the same as greater benefits for citizens or other actors. Evaluation must consider effects on service recipients and others, as well as costs within the organization.

5.4 Clients and AI Professionals: Delegating Judgment and the Difficulty of Verification

AI acting as a professional is an AI agent entrusted with a defined part of diagnosis, assessment of options, recommendations, or implementation in matters about which the client lacks sufficient knowledge. Clients may be able to communicate the results they want without being able to judge what is necessary to achieve them or which means are

appropriate. They may have to rely on professional judgment both in specifying what to request and in evaluating the results.

In the credence goods analyzed by Dulleck and Kerschbamer (2006), users cannot sufficiently ascertain the level of service they need and may find it difficult, even after a problem has been resolved, to judge whether they received more service than necessary. What was needed, what was actually done, and what was charged for are separate questions. Their basic model assumes, however, that experts can diagnose accurately and maximize profits. This paper makes no such assumption about profit motivation for AI, nor can it assume that AI always diagnoses correctly.

Possible causes of erroneous AI recommendations include limits to knowledge or reasoning ability, ambiguity in the request, and inappropriate evaluation criteria. The information problem that a user cannot detect an error must be distinguished from the motivational problem of an expert concealing information for private gain. This paper evaluates limitations in AI capability themselves and examines the losses arising when delegation disregards known limitations or necessary verification and consultation are omitted.

It is also important to distinguish advice from execution. Crawford and Sobel (1982) analyze a relationship in which a sender with private information communicates it and the receiver chooses the final action. This is a theory of strategic communication by an informed party among actors with divergent interests, rather than a direct explanation of why AI generates incorrect content. This paper examines separately how far judgment in formulating advice and execution of actions based on that advice are entrusted to AI. Discretion over professional judgment can be delegated even without granting authority to act externally. Nor does a procedure for final human approval necessarily imply adequate oversight.

The framework requires distinguishing value judgments that clients themselves should make from matters requiring professional expertise. For example, even when AI assesses whether a method is technically appropriate, the acceptable costs, time, and inconvenience depend on the client's preferences. It is important to present the grounds for a recommendation, conditions for its applicability, uncertainties, and alternatives so that clients can decide or request further verification.

Simply increasing the number of verifiers is not sufficient, however. Agreement between two AI systems using the same information and sharing the same weaknesses in reasoning does not by itself constitute independent corroboration. Analysis must consider whether additional knowledge or evidence is available, whether the verifier has authority to correct errors, and whether improvements justify verification costs. If a provider or service seller is entrusted with setting diagnostic methods or recommendation rankings and clients cannot fully observe those settings, incentives to increase referral fees or sales may create a separate agency problem. This is treated distinctly from private utility of AI itself.

5.5 Capital Providers and AI Fund Managers: Specifying Risk and Evaluating Performance

AI acting as a fund manager is an AI agent entrusted with a defined range of asset allocation, security selection, trade execution, or related tasks in accordance with capital providers' objectives. The relationship with LP investors mentioned in the introduction may involve LPs entrusting funds to an investment manager, which in turn redelegates some tasks to AI. Direct delegation from capital providers to AI must be distinguished from delegation through an asset-management firm or staff member, clarifying who sets investment policy and who oversees its implementation.

For human asset management, Chevalier and Ellison (1997) analyze mutual funds to examine the relationship between the investment performance sought by investors and the benefits fund companies obtain through fund inflows. They estimate the incentives facing fund companies from the relationship between past performance and inflows and report that actual changes in risk are consistent with those incentives. This study informs the analysis of investment managers' incentives but does not directly test LP-type funds or AI behavior.

When investment management is delegated to AI, returns alone do not adequately specify the objective. Conditions such as the investment horizon, tolerance for losses, when funds must be liquidated, and permitted transactions are also important. This paper does not assume that AI itself takes risks in pursuit of management fees or reputation. If AI is given only a short-term return target, however, it may choose actions that inadequately account for long-term capital preservation or liquidity. Even without

assuming private risk preferences for AI itself, the risks borne by capital providers must be managed.

Evaluating performance requires separating realized returns from the appropriateness of delegated judgments. Losses caused by broad market movements do not alone imply AI agency slack. Conversely, high returns may have been obtained through investment concentration or borrowing beyond permitted limits. Evaluation requires a benchmark suited to investment objectives, along with examination of information available at the time of trading, risks undertaken, and compliance with constraints. Rather than using the investment that would have produced the highest return in hindsight, the assessment should examine which judgments were desirable given the information and feasible choices at the time.

Under this framework, conditions concerning eligible assets, concentration, borrowing, transaction size, and other factors must be defined, and AI execution authority and verification methods designed accordingly. The allocation of computing resources is also distinct from the allocation of investment capital or the setting of permissible risk. More computation enabling consideration of more options does not authorize large transactions. Nor does additional computation necessarily improve judgment; its benefits must be compared with its costs and time requirements.

In considering interruptibility, it is necessary to distinguish stopping new decisions or orders, cancelling unexecuted orders, and disposing of existing holdings. Stopping AI leaves the risk of existing holdings in place. Selling assets also incurs transaction costs and may itself affect prices. Whether to interrupt must therefore be judged by comparing the benefits and losses of allowing AI to continue managing investments with those of managing assets through feasible alternatives after stopping it. Delegation must be designed to specify who decides to stop execution and who takes over subsequent management of orders and assets.

5.6 Comparing Delegation Relationships

What the four relationships share is that even without assuming AI preferences over compensation, behavior aligned with the principal's objectives is not guaranteed and is not easy to evaluate. The contexts in which problems arise nevertheless differ. The preceding discussion can be summarized as follows.

Role	Issues in defining objectives and delegated scope	Difficulties in evaluation	Distinctions needed for intervention
AI worker	Task changes, priorities, and division of work with others	Processing volume and speed alone do not capture the quality of the overall workflow	Stopping tasks, handing over unresolved cases, and effects on other departments
AI bureaucrat	Priorities among objectives; authority to change criteria and process individual cases	Losses differ by error type; some problems do not appear in objections	Stopping processing, reassessment, correcting erroneous processing, and remedying harm
AI professional	Client preferences, professional judgment, and the scope of implementation authority	Clients struggle to judge whether recommended means are necessary and appropriate	Correcting advice, stopping implementation, and further verification
AI fund manager	Investment horizon, liquidity, permissible risk, and trading authority	Separating market effects from the appropriateness of judgments at the time of trading	Stopping order placement, cancelling orders, and managing or liquidating holdings

The problems in the table are not confined to their respective domains. Firms also have multiple objectives, and difficulties in verifying professional judgment apply to public administration and asset management as well. What matters is to establish, for each delegation relationship, who can determine objectives, who can evaluate the appropriateness of behavior and when, and what can be corrected or recovered after interruption. Even with the same AI model, the problems of delegation change with the authority granted and the way AI is incorporated into the work.

Costs are compared using the classification from Section 3:

$$AC_{AI} = S + M + I + R \tag{3}$$

Here, (S) denotes the costs of drafting and updating objectives, constraints, and evaluation criteria; (M), monitoring and verification costs; (I), costs directly required to control authority, approve actions, modify execution, interrupt, or transfer tasks; and (R), the expected loss that remains. Which costs become large, however, is not determined solely by the name of the domain.

For example, without agreement on priorities among objectives, stronger monitoring still leaves undefined what counts as improvement. Even if the appropriateness of behavior can be assessed, that information is difficult to use for correction if stopping, modification, or transfer is not possible when needed. Even when interruption is possible, it cannot alone prevent losses after the fact if consequences already incurred cannot be reversed. Rather than minimizing each cost separately, specification, observation, authority setting, and intervention must therefore be considered in combination.

Cost accounting must also avoid overlap among components. The personnel and time costs of approval are distinguished from benefits forgone while awaiting approval and losses from mistaken interruption; the latter are treated as residual loss. Nor is it appropriate to include all computing costs required for ordinary task execution in agency costs. Agency costs attributable to human monitors or providers, and losses to third parties outside the delegation relationship, also require evaluation that clarifies their relationship to the costs of delegation to AI.

5.7 Implications of This Section

The comparison shows that the effectiveness of delegation to AI cannot be judged solely from AI capability or usage charges. Analysis must include the costs of specifying objectives, evaluating behavior, and intervening as necessary, together with the losses that remain despite these measures. It should compare realistically feasible arrangements for human performance, AI delegation, and a division of work between them, rather than assuming perfect execution.

Expanding delegation gives AI greater scope to respond to circumstances. It may also increase the occasions on which AI effectively determines unspecified priorities among objectives or methods of implementation. Conversely, narrower authority restricts behavior but may sacrifice benefits through waiting for approval or frequent handovers. The desirable extent of delegation depends on the verifiability of judgments, the consequences of failure, the possibility of correcting or recovering outcomes, and the capability of the supervising party. This section's analysis alone cannot establish an optimal scope of delegation common to all domains.

What is retained from conventional agency theory is the perspective of recognizing the benefits of entrusting information gathering and decision-making to others while

analyzing the associated losses. This question persists even when AI is not assumed to have independent private utility. The mechanism through which governance affects AI behavior, however, differs from that for humans. Control of the scope of delegation, allocation of computing resources, and interruptibility are understood not as rewards that motivate AI, but as arrangements determining what AI is entrusted with, under what conditions it operates, and how it can be corrected or stopped when necessary.

6 Conclusion

This paper defined a principal–AI agent relationship as a relationship in which a human or organization entrusts AI with performing tasks and a degree of discretion over judgment or execution, and presented a framework for analyzing the problems arising within it. The baseline setting assumes neither independent private utility nor preferences over compensation for AI itself. Delegation can nevertheless fail if there are problems in communicating objectives, evaluating behavior, or making necessary corrections and interruptions. The absence of AI's pursuit of private interests is not the same as behavior aligned with the principal's objectives.

To capture this problem, the paper called divergence between the principal's intentions and the AI's behavior arising from constraints on delegation AI agency slack. It classified its sources into underspecification, misspecification, and noncompliance with specifications, distinguishing objective specification, behavior, and outcomes. It also classified the costs of addressing divergence and the losses remaining afterward as AI agency costs. This provides a perspective for evaluating the drafting and updating of objectives and constraints, monitoring and verification, control of authority and intervention, and residual loss together with the benefits of delegation. Poor outcomes alone do not justify treating losses from general capability limitations or the external environment as agency problems.

Ante (2026) focuses on losses arising from the gap between the principal's intentions and the objective given to AI, even in the absence of conflicts of interest. Building on this insight, this paper situated the evaluation of judgments, allocation of authority, and feasibility of intervention alongside objective specification within an analytical framework based on existing contract theory. Conventional theory has addressed not only compen-

sation but also information acquisition, the division of tasks, decision control, and the allocation of authority. The contribution of this paper is to retain these analyses while clarifying which problems remain and how governance operates when AI is not assumed to have independent private utility.

Under this baseline assumption, governance of AI cannot be understood as changing motivation through compensation. What the principal designs are the conditions of delegation: what objective to communicate, how far judgment and execution are entrusted to AI, how much computing capacity to allocate, and under what conditions intervention is possible. Computing resources are inputs for execution and verification, rather than rewards for AI. Interruptibility likewise is not a means of deterring behavior by threatening AI with adverse consequences, but a mechanism for stopping processing when necessary and enabling correction or handover. This requires not only formal authority but also the information needed for judgment, the time to intervene, and actual means of changing processing or external operations.

In applying the framework to the shareholder–AI manager relationship, the paper examined the design of the decision process: how shareholders' objectives are reflected in managerial instructions and who undertakes initiation, ratification, implementation, and oversight. It further compared AI in the roles of worker, bureaucrat, professional, and fund manager, clarifying how governance challenges differ with who determines objectives, how well judgments can be evaluated, the losses caused by errors, and the possibility of correcting or recovering outcomes after intervention. Organizations using AI also include humans and organizations with their own interests, such as monitors and providers. Not assuming private interests for AI itself does not justify concluding that conflicts of interest surrounding delegation disappear as well.

Neither expanding delegation to AI nor strengthening oversight can therefore be regarded as desirable in itself. Broader discretion allows benefits from judgments responsive to circumstances, but may also increase the consequences of divergence from intentions. Approval and restrictions on authority may reduce losses but create checking burdens and delays. Comparing human performance, delegation to AI, and a division of work between them requires jointly examining the benefits of execution, ordinary execution costs, governance costs, and residual loss for the same tasks and evaluation criteria.

What this paper presented is a conceptual framework for making such comparisons. It did not empirically establish the effects of individual governance instruments or an optimal scope of delegation. Future research should specify in advance the objectives and constraints important to the principal and test how behavioral divergence and costs change as the conditions of instructions, verification, authority, and intervention vary. Differences in AI capability and available information must be considered and distinguished from the effects of delegation design. How to aggregate principals' objectives, how to evaluate losses to third parties, and how to extend the framework when AI should be assumed to possess objectives of its own also remain questions for future research.

As delegation to AI expands, the question is not only how much work AI can perform, but whose objectives its judgments should serve, who evaluates them, and who can change them when necessary. Governance supporting delegation remains necessary even without assuming private interests for AI. This paper's framework provides a foundation for considering AI capability together with the conditions for its appropriate use within organizations.

References

Renee B. Adams, Benjamin E. Hermalin, and Michael S. Weisbach. The role of boards of directors in corporate governance: A conceptual framework and survey. *Journal of Economic Literature*, 48(1):58–107, 2010. doi: 10.1257/jel.48.1.58.

Philippe Aghion and Jean Tirole. Formal and real authority in organizations. *Journal of Political Economy*, 105(1):1–29, 1997. doi: 10.1086/262063.

Ricardo Alonso and Niko Matouschek. Optimal delegation. *The Review of Economic Studies*, 75(1):259–293, 2008. doi: 10.1111/j.1467-937X.2007.00471.x.

Lennart Ante. Specification cost: Relocating the binding constraint in delegation to AI agents with externally supplied objectives. *Technology in Society*, 88:103503, 2026. doi: 10.1016/j.techsoc.2026.103503.

Susan C. Athey, Kevin A. Bryan, and Joshua S. Gans. The allocation of decision authority to human and artificial intelligence. *AEA Papers and Proceedings*, 110:80–84, 2020. doi: 10.1257/pandp.20201034.

George Baker, Robert Gibbons, and Kevin J. Murphy. Subjective performance measures in optimal incentive contracts. *The Quarterly Journal of Economics*, 109(4):1125–1156, 1994. doi: 10.2307/2118358.

George P. Baker. Incentive contracts and performance measurement. *Journal of Political Economy*, 100(3):598–614, 1992. doi: 10.1086/261831.

Marianne Bertrand and Sendhil Mullainathan. Are CEOs rewarded for luck? the ones without principals are. *The Quarterly Journal of Economics*, 116(3):901–932, 2001. doi: 10.1162/00335530152466269.

Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. Generative AI at work. *The Quarterly Journal of Economics*, 140(2):889–942, 2025. doi: 10.1093/qje/qjae044.

Judith Chevalier and Glenn Ellison. Risk taking by mutual funds as a response to incentives. *Journal of Political Economy*, 105(6):1167–1200, 1997. doi: 10.1086/516389.

Vincent P. Crawford and Joel Sobel. Strategic information transmission. *Econometrica*, 50(6):1431–1451, 1982. doi: 10.2307/1913390.

James H. Davis, F. David Schoorman, and Lex Donaldson. Toward a stewardship theory of management. *Academy of Management Review*, 22(1):20–47, 1997. doi: 10.5465/amr.1997.9707180258.

Wouter Dessein. Authority and communication in organizations. *The Review of Economic Studies*, 69(4):811–838, 2002. doi: 10.1111/1467-937X.00227.

Mathias Dewatripont and Jean Tirole. Advocates. *Journal of Political Economy*, 107(1):1–39, 1999. doi: 10.1086/250049.

Avinash Dixit. Incentives and organizations in the public sector: An interpretative review. *The Journal of Human Resources*, 37(4):696–727, 2002. doi: 10.2307/3069614.

Uwe Dulleck and Rudolf Kerschbamer. On doctors, mechanics, and computer specialists: The economics of credence goods. *Journal of Economic Literature*, 44(1):5–42, 2006. doi: 10.1257/002205106776162717.

Kathleen M. Eisenhardt. Agency theory: An assessment and review. *Academy of Management Review*, 14(1):57–74, 1989. doi: 10.5465/amr.1989.4279003.

Eugene F. Fama. Agency problems and the theory of the firm. *Journal of Political Economy*, 88(2):288–307, 1980. doi: 10.1086/260866.

Eugene F. Fama and Michael C. Jensen. Separation of ownership and control. *The Journal of Law and Economics*, 26(2):301–325, 1983. doi: 10.1086/467037.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10835–10866. PMLR, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Luis Garicano. Hierarchies and the organization of knowledge in production. *Journal of Political Economy*, 108(5):874–904, 2000. doi: 10.1086/317671.

Sanford J. Grossman and Oliver D. Hart. The costs and benefits of ownership: A theory of vertical and lateral integration. *Journal of Political Economy*, 94(4):691–719, 1986. doi: 10.1086/261404.

Dylan Hadfield-Menell and Gillian K. Hadfield. Incomplete contracting and AI alignment. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 417–422. Association for Computing Machinery, 2019. doi: 10.1145/3306618.3314250.

Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. The off-switch game. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 220–227. International Joint Conferences on Artificial Intelligence Organization, 2017. doi: 10.24963/ijcai.2017/32.

Oliver Hart and Luigi Zingales. Companies should maximize shareholder welfare not market value. *Journal of Law, Finance, and Accounting*, 2(2):247–275, 2017. doi: 10.1561/108.00000022.

Oliver Hart and Luigi Zingales. The new corporate governance. *The University of Chicago Business Law Review*, 1(1):195–216, 2022. URL https://businesslawreview.uchicago.edu/print-archive/new-corporate-governance.

Oliver Hart, Andrei Shleifer, and Robert W. Vishny. The proper scope of government: Theory and an application to prisons. *The Quarterly Journal of Economics*, 112(4):1127–1161, 1997. doi: 10.1162/003355300555448.

Bengt Holmström. Moral hazard and observability. *The Bell Journal of Economics*, 10 (1):74–91, 1979. URL https://www.jstor.org/stable/3003320.

Bengt Holmström. Pay for performance and beyond. *American Economic Review*, 107 (7):1753–1777, 2017. doi: 10.1257/aer.107.7.1753.

Bengt Holmström and Paul Milgrom. Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *The Journal of Law, Economics, and Organization*, 7(Special Issue):24–52, 1991. doi: 10.1093/jleo/7.special_issue.24.

Bengt Holmström and Paul Milgrom. The firm as an incentive system. *American Economic Review*, 84(4):972–991, 1994. URL https://www.jstor.org/stable/2118041.

Beth K. Humberd and Scott F. Latham. When AI becomes an agent of the firm: Examining the evolution of AI in organizations through an agency theory lens. *Journal of Management Studies*, 63(2):668–694, 2026. doi: 10.1111/joms.13274.

Michael C. Jensen. Agency costs of free cash flow, corporate finance, and takeovers. *American Economic Review*, 76(2):323–329, 1986. URL https://www.jstor.org/stable/1818789.

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

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293, 2018. doi: 10.1093/qje/qjx032.

Meta. Introducing Muse: The world's first personal AI agent built for everyone. Meta Newsroom, 2026. URL https://about.fb.com/news/2026/09/introducing-muse-personal-ai-agent/. September 8, 2026; page updated September 30, 2026. Accessed November 1, 2026.

OpenAI. Introducing dots. Product announcement, 2026. URL https://openai.com/index/introducing-dots/. September 29, 2026. Accessed November 1, 2026.

Canice Prendergast. The motivation and bias of bureaucrats. *American Economic Review*, 97(1):180–196, 2007. doi: 10.1257/aer.97.1.180.

Stephen A. Ross. The economic theory of agency: The principal's problem. *American Economic Review*, 63(2):134–139, 1973. URL https://www.aeaweb.org/aer/top20/63.2.134-139.pdf.

Andrei Shleifer and Robert W. Vishny. A survey of corporate governance. *The Journal of Finance*, 52(2):737–783, 1997. doi: 10.1111/j.1540-6261.1997.tb04820.x.

Herbert A. Simon. A formal theory of the employment relationship. *Econometrica*, 19 (3):293–305, 1951. doi: 10.2307/1906815.

Michael Wooldridge and Nicholas R. Jennings. Intelligent agents: Theory and practice. *The Knowledge Engineering Review*, 10(2):115–152, 1995. doi: 10.1017/S0269888900008122.