Instruction Theory: Complete and Incomplete Instructions and the Boundary Between Humans and AI

Seiryu Ando

Graduate School of Economics, Kyoto University

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

November 3, 2026 (Theoretical Draft)

Abstract

This paper analyzes delegation from humans to AI by asking what should be specified in advance, which decisions should remain with humans, and which tasks should be assigned to AI. Building on the analysis of ex ante specification, authority allocation, and information transmission in contract theory and organizational economics, it formulates these issues as a problem of instruction design. An instruction is defined as complete if the candidate intents consistent with the instruction and shared context imply the same set of permissible actions in every relevant situation, and as incomplete otherwise. Even under an incomplete instruction, missing information about intent causes no loss if a common optimal action exists. A model with a binary decision criterion compares direct performance by a human, delegation to AI under complete instructions, and delegation to AI combining incomplete instructions with consultation on individual cases. A common attention constraint on human execution, instruction preparation, and responses to consultations jointly determines task assignment and the division of decision making within each task. Holding fixed an incomplete-instruction regime for a given task type, a higher shadow price of attention weakly reduces consultation, but need not induce a shift toward complete instructions because their preparation also requires time. The paper further analyzes a setting in which the same criterion persists and answers can be remembered, deriving the future value of consultation and the choice of ex ante specification. Using general principles of the value of information, it presents a limited theoretical model that makes human involvement through instruction, consultation, and execution comparable.

Keywords: instruction theory, complete instructions, incomplete instructions, AI delegation, value of information, human attention

JEL classification: D83, D23, D86, O33

1 Introduction

Delegation to AI agents is expanding from information collection and analysis to task execution and decision making. Humans communicate objectives and delegate to AI the selection and execution of the actions required to achieve them. The question is therefore not only what AI can execute, but also what humans should specify in advance and what they should decide after a case arises. Unspecified preferences in user requests and the burden that further clarification imposes on users have also become direct objects of analysis in research on collaboration with AI (Y. R. Dong et al., 2026).

This paper uses the term instruction theory for a framework concerned with the content of instructions and the allocation of decisions that this content alone does not determine. It seeks to apply and give concrete form to the analysis of ex ante specification and ex post decision making in contract theory in the context of delegation to AI. The question of whether task content should be determined in advance or left to later judgment was already explicit in Simon (1951). Comparing the cost of possessing information with the cost of asking others when it is needed is also central to Garicano (2000). Furthermore, Athey, Bryan, and Gans (2020) analyze the allocation of decision authority to humans and AI, including its effects on human information-acquisition effort. Building on these contributions, this paper examines the division of work through the transmission of the principal's decision criterion and subsequent consultation.

Even when AI can correctly recognize the facts relevant to a task, what the principal wants may not have been sufficiently communicated. For example, an AI that accurately compares prices and quality may still be unable to infer from a request how much deterioration in quality is acceptable in exchange for lower costs. Additional computation alone need not resolve this deficiency. On the other hand, decisions may coincide for cases of sufficiently high quality or clearly excessive cost even when the principal's precise criterion is unknown. It is therefore necessary to distinguish an instruction being incomplete from human confirmation being required.

To make this distinction, the paper defines completeness in terms of the sets of permissible actions implied by candidate intents consistent with the instruction and shared context. If these sets coincide in every relevant situation, the instruction is complete. This definition does not require a single action to be prescribed. It is enough that the scope of discretion accepted by the principal be determined. Moreover, even when the

permissible-action sets differ, losses can be avoided without additional information if there is an optimal action common to every candidate intent. Content completeness, sufficiency of information for decision making, and the ability to execute actions are distinct properties.

Under incomplete instructions, consultation is determined by comparing the reduction in expected loss due to the answer with the cost of obtaining it. This condition is a general principle of the value of information, and existing research also addresses the selection of questions by AI. In addition to the consultation policy, this paper considers the preceding stage of task design. The principal can perform a task personally, communicate a criterion in advance and delegate the task to AI, or leave part of the criterion unspecified and provide answers when needed. Each option uses the principal's time. Consequently, deciding whether to ask a question in a particular case is not sufficient to determine the division of work between humans and AI.

This paper solves this choice explicitly using a binary decision criterion and a continuous case evaluation. First, for a fixed instruction, it derives regions in which AI autonomously declines to act, consults the human, or autonomously acts. It then adds direct human performance as an alternative and imposes a common attention constraint on instruction preparation and responses to consultations. This distinguishes and jointly determines two boundaries: which tasks are delegated to AI, and which cases within delegated tasks are referred back to humans.

The analysis yields three main implications. First, even when complete instructions are feasible, incomplete instructions combined with selective consultation can be less costly. Second, when human attention becomes scarcer, consultation weakly decreases under a given instruction, but whether the adoption of complete instructions increases depends on the time required to prepare them. Third, when answers can be reused in later cases, consultation has future informational value, and its comparison with ex ante specification cannot be obtained simply by multiplying the one-shot cost by the number of cases.

The remainder of the paper reviews the related literature and defines instruction completeness. It then analyzes action selection for individual cases, task assignment and attention allocation, and repeated delegation with reusable answers. Finally, it discusses testable implications and the model's limitations. The analysis identifies optimal designs under specified instruction technologies and information structures; it does not characterize a general optimal mechanism encompassing all linguistic expressions or contracts.

2 Related Literature and the Position of This Paper

2.1 Ex Ante Specification and Ex Post Judgment

Coase (1937, §I and the beginning of §II) treats coordination through prices and direction within the firm as alternative methods of coordination, emphasizing the cost of using the price mechanism. This paper adopts the perspective that task assignment and coordination methods should be compared in terms of their costs of use, rather than determined solely by technical feasibility. Extending this perspective to the cost of instructions to AI is an application made in the present paper.

Simon (1951) formalizes both contracts that specify behavior in advance and employment relationships that confer authority to select behavior later within a zone of acceptance. Section 7, in particular, presents a structure that allocates elements of a task among advance specification, the employer's judgment, and the worker's discretion. Instead of transferring directly to AI a mechanism based on workers' private utility and wage bargaining, the present paper explicitly models the cost of communicating the principal's criterion and the resources needed for responses and execution.

Grossman and Hart (1986, §I) understand ownership in terms of residual rights of control over uses of assets that are not individually specified in a contract. This paper likewise specifies who decides on unspecified matters, but its object is decision authority within tasks. It examines the use of information through such authority.

2.2 Information, Cognition, and Authority

Aghion and Tirole (1997, §§I-IV) distinguish formal decision rights from authority that actually influences decisions through information. This distinction provides a basis for understanding why retaining approval rights for a human does not by itself guarantee effective judgment. Their incentives for information acquisition, however, depend on the agents' private benefits. The cooperative AI in the present paper is treated as following a separately specified decision rule.

Garicano (2000, §§II-IV) explains the division of problem-solving work through the cost of acquiring knowledge and the time cost of seeking solutions from others. Evaluating consultation with the principal in terms of informational value and time cost is common to that study and the present one. Here, however, the information possessed by the human

is primarily the principal's own evaluation criterion, rather than expertise concerning external facts. The paper compares ex ante communication, case-specific consultation, and direct human performance under the same cost and time constraints, and derives their concrete boundaries.

Tirole (2009) analyzes the costly process of recognizing design problems and incorporating the resulting information into contracts. Sections II and IV of the author's draft dated March 25, 2008, consulted for this paper, also establish the importance of treating contractual detail as an endogenous choice. In the present paper, the principal's criterion is already determined, and the timing and method of its communication are chosen. The cost of identifying previously undiscovered design flaws is distinguished from the cost of communicating a known criterion.

2.3 Incompleteness and Implementability

Contract theory is not restricted to settings in which all actions are observable and drafting contracts is costless. Holmström (1979), for example, studies incentives and risk sharing under unobservable actions. This paper abstracts from incentive problems and focuses on informational constraints concerning intent.

Nor does the inability to describe content in advance alone imply that a desirable outcome cannot be implemented. Maskin and Tirole (1999, §§3-6) and Tirole (1999) examine implementability, including ex post information exchange and mechanisms. The present paper likewise does not infer loss or inefficiency directly from incomplete instructions. It includes both cases in which decisions coincide without additional information and cases in which consultation can resolve the deficiency, and specifies the conditions under which such supplementation is costly.

2.4 The Allocation of Decision Authority to Humans and AI

Athey, Bryan, and Gans (2020, §§I-II) analyze final decision authority and human information-acquisition effort in a setting with a principal, a human agent, and AI. In their basic model, an informed AI selects the principal's preferred project, whereas the human agent has distinct preferences and incurs effort costs. Retaining authority with the human is evaluated not only in terms of agreement with the principal's judgment

but also in terms of incentives to acquire information. This study is directly relevant to the allocation of decisions between humans and AI examined here.

The human who directly performs tasks or answers consultations in this paper is the principal who issues the instruction. Accordingly, regime H in Section 6 is not the same as the human agent H in that study. Here, AI observes the facts of the case x accurately, while the principal knows the decision criterion θ, but communicating that criterion to AI is costly. The question is when and by what method known information should be made available; human-agent information-acquisition effort is not endogenized.

Moreover, in that study a human can use AI recommendations even while holding decision authority. The location of final decision authority therefore cannot be equated with responsibility for performing an entire task. This paper separates the choice among three regimes of task assignment from the choice of case-specific consultation after delegation. By making the agents and information structure explicit, it relates the analysis of how decision rights change incentives for information acquisition to the analysis of how instructions and consultation change the cost of using the principal's information.

2.5 Instructions to AI and Communication

Hadfield-Menell and Hadfield (2019) connect the specification of AI objectives with incomplete contracts and discuss supplementation through law and social norms. Ante (2026, §§3-5) organizes the design and verification of objectives and residual losses as specification costs for AI whose objectives are externally assigned. The persistence of insufficient communication of intent even without conflicting interests, and the supplementation of specifications through external procedures, have already been discussed in these studies.

J. Dong, Jhunjhunwala, and Kanoria (2026) jointly consider the cost of communicating preferences to cooperative AI and the cost to humans of choosing among proposed alternatives. Y. R. Dong et al. (2026, §4.1) express the benefit of asking about unspecified intent as a value of information and compare it with the cost of the question. The condition for a single consultation in this paper shares this general logic. Fudenberg and Liang (2026) study the design of delegation and restrictions on input information when AI alignment is uncertain. This paper assumes no uncertainty about alignment and examines communication costs that remain even under a cooperative decision rule.

The paper distinguishes instruction completeness from the information needed for decision making and connects ex ante specification, consultation, and human performance within a single resource-allocation problem. In particular, its static and dynamic models make explicit, respectively, that instruction preparation requires human time and that answers can be reused in subsequent cases.

3 Instructions and Their Completeness

3.1 Candidate Intents and Shared Context

The human delegating a task, or the authorized decision maker of an organization doing so, is called the principal. Let the relevant situation be $s \in S$, the feasible set of actions be $A(s)$, and intent, encompassing the principal's objectives, constraints, and priorities, be θ.

Definition 1 (Instruction). An instruction I specifies objectives, constraints, priorities, and scope of application communicated by the principal to guide AI's action selection. Let K_0 denote shared context, such as referenced rules, definitions of terms, and prior agreements, and let $\Theta(I, K_0)$ denote the nonempty set of candidate intents consistent with these.

In the basic analysis, the principal's true intent belongs to this set. Even clear wording constitutes misspecification if it excludes the true intent; this is distinguished from incompleteness. Shared context is fixed before the analysis, and content inferred subsequently by AI is not counted as information agreed upon from the outset.

Let $D_\theta(s) \subseteq A(s)$ be the nonempty set of actions the principal permits in situation s under intent θ.

Definition 2 (Complete Instruction). An instruction satisfying the following condition is complete in content with respect to the relevant situations S and shared context K_0:

$$D_\theta(s) = D_{\theta'}(s) \quad \text{for all } s \in S,\ \theta, \theta' \in \Theta(I, K_0). \tag{1}$$

Definition 3 (Incomplete Instruction). An instruction for which equation (1) does not hold is called incomplete.

Completeness does not require identifying every aspect of intent. It suffices that the actions permitted in the relevant task be the same. The permissible set may also contain

multiple actions. If the scope of discretion is established, as in "an order may be placed with any of the three firms that meet the conditions," there is no incompleteness in that respect. Conversely, even a long instruction can be incomplete if priorities in exceptional circumstances remain unspecified.

3.2 Information Sufficient for Decision Making

Content completeness is not necessary for selecting an appropriate action. For example, if the permissible sets under two intents are $\{a, b\}$ and $\{b, c\}$, respectively, the sets differ, but action b conforms to both.

Definition 4 (Decision Sufficiency). An instruction is sufficient for selecting a conforming action in situation s if the permissible-action sets of the candidate intents have a nonempty intersection:

$$\bigcap_{\theta \in \Theta(I, K_0)} D_\theta(s) \neq \emptyset. \tag{2}$$

This definition separately requires that selection from the intersection be informationally and computationally feasible. In general, permissible sets may include actions of different value to the principal. In the cost and loss analysis in the following sections, permissible sets are specified as the principal's sets of optimal actions. Decision sufficiency can then be linked to the condition for zero loss from missing information about intent.

3.3 Supplementary Procedures and Decision Authority

Even when content is incomplete, it is possible to specify how missing information should be supplied. One example is a procedure that consults the principal when an unspecified condition affects the decision and suspends execution until an answer arrives.

Definition 5 (Procedural Completeness). A supplementary procedure is complete if, for the relevant observation histories, it determines who supplies the missing information, the conditions for consultation, what happens when no answer is received, and who is responsible for revision and stopping.

The operational authority to resolve unspecified matters and determine actions is called residual decision authority. It is distinguished from authority to investigate, propose, or execute external operations. For example, AI may place an order after a human has

decided whether a case should be accepted. Human responsibility for a decision does not imply that the human performs the entire task again.

The following analysis assumes that procedures are reliably executed, that the information needed for consultation reaches the principal, and that execution can wait for an answer. Retaining formal approval authority alone in an arrangement that does not satisfy these conditions will not deliver the informational value described by the model.

4 Loss from Missing Information and Consultation

4.1 Cooperative Action Selection

Let the utility of an action be $U(a, s, \theta)$, and define the value and set of optimal actions when the principal's criterion is known as

$$V^*(s, \theta) = \max_{a \in A(s)} U(a, s, \theta), \qquad D_\theta(s) = \arg\max_{a \in A(s)} U(a, s, \theta) \tag{3}$$

The decision makers compared here use the same factual information and action technology. U is expected utility at the time of the decision; the benchmark does not have exclusive knowledge of future outcomes.

Let h be a history including the instruction, shared context, and observed situation, and let $p(\theta \mid h)$ be the conditional distribution of the principal's intent. AI is assumed to maximize the principal's expected utility under this distribution. This is a cooperative decision rule imposed for the analysis; it does not follow solely from the absence of private interests.

The minimum expected loss without additional information is

$$r(h) = \mathbb{E}[V^*(s, \theta) \mid h] - \max_a \mathbb{E}[U(a, s, \theta) \mid h] \geq 0 \tag{4}$$

This is the loss due to not knowing the principal's criterion, and is distinguished from errors in recognizing facts or executing operations.

Proposition 1 (When Missing Information Causes Loss). Suppose that, conditional on a fixed history, the candidate actions and intents are finite and each candidate intent has strictly positive conditional probability. Then

$$r(h) = 0 \quad \Longleftrightarrow \quad \bigcap_{\theta \in \operatorname{supp} p(\cdot|h)} D_\theta(s) \neq \emptyset. \tag{5}$$

If a common optimal action exists, loss can be avoided without identifying intent. If the intersection is empty, every action leaves a strictly positive expected loss. As long as AI already performs the optimization in equation (4), faster computation using the same information alone cannot eliminate this loss. When additional history or consultation supplies new information, the conditional distribution itself changes.

Completeness is a property concerning every relevant situation. By contrast, zero expected loss requires a condition only on intents with positive probability and on histories that occur almost surely. These two concepts should not be equated.

4.2 The Informational Value of Consultation

Let Y be the answer obtained from a single consultation, and let $r^Q(h)$ be the expected loss from optimal behavior after the answer, evaluated using the information available before the answer. Let the cost of consultation be $c(h)$.

Lemma 1 (Consultation Based on the Value of Information). When only the current case is considered and there is no aggregate capacity constraint, consultation is strictly desirable if and only if

$$r(h) - r^Q(h) > c(h) \tag{6}$$

At equality, both policies are optimal.

This condition applies the principle of the value of information and shares the logic of the evaluation of questions in Y. R. Dong et al. (2026, §4.1). If answers contain errors or reveal only part of the information, $r^Q(h)$ need not be zero. Moreover, when multiple questions are complementary or answers can be reused later, the informational value for the current case alone does not determine the optimal sequential policy.

4.3 Instruction Technologies and the Scope of the Analysis

In general, instruction design is the choice of an information structure from among available methods of communication. Let e be a communication technology selected ex ante, k_e its cost, and m the message generated as a function of the principal's criterion. Once the posterior distribution induced by the message and the consultation methods available thereafter are specified, the instruction cost can be compared with the expected costs of consultation and residual loss.

This paper does not, however, solve a general theory that maps every expression in natural language into a cost and an information structure. Sections 5 and 6 use a regime that communicates the criterion exactly and a regime that leaves it uncommunicated and instead permits consultation on individual cases. The preparation cost of the former includes not only sending a code but also documenting and checking the criterion and configuring it in the execution environment. If partial communication, the provision of examples, or the retrieval of external rules is more advantageous, the problem must be solved again over an expanded set of instruction technologies that includes these options.

This restriction neither assumes that complete instructions generally minimize cost nor permits only incomplete instructions. It specifies the available regimes so that optimality can be examined within that set.

5 Within-Task Decision Boundaries

5.1 Case Evaluation and the Principal's Criterion

Let $a = 1$ denote implementing a case and $a = 0$ denote rejecting it. The AI observes the case evaluation $x \in [0, 1]$ accurately, and the principal's utility is

$$U(a, x, \theta) = Da(x - \theta), \qquad D > 0 \tag{7}$$

Here, D represents the importance of the decision, and θ is the threshold for implementation. If the threshold is known, the case is implemented when $x > \theta$ and rejected when $x < \theta$.

The threshold takes one of two values, $0 < \theta_L < \theta_H < 1$, with $\Pr(\theta = \theta_H) = q \in (0, 1)$. When the case is processed, the principal knows the realized threshold, but the AI does

not know its realization if it has not been specified. The regime considered in Section 6 is selected before this threshold is realized. The case evaluation x is uniformly distributed and independent of the threshold. Define

$$\mu = (1-q)\theta_L + q\theta_H, \quad \Delta = \theta_H - \theta_L, \quad \kappa = Dq(1-q) \tag{8}$$

To maximize expected utility, an AI that does not yet know the principal's threshold implements the case when $x \geq \mu$ and rejects it otherwise. At $x = \mu$, the AI is indifferent; we adopt the rule that it implements the case.

5.2 The Value of Knowing Intent

The expected loss relative to knowing the principal's threshold is

$$r(x) = \begin{cases} 0, & x \leq \theta_L, \\ D(1-q)(x-\theta_L), & \theta_L < x < \mu, \\ Dq(\theta_H - x), & \mu \leq x < \theta_H, \\ 0, & x \geq \theta_H. \end{cases} \tag{9}$$

This loss includes both implementing an unsuitable case and rejecting a suitable one. Different thresholds lead to different implementation decisions in the region (θ_L, θ_H). Outside this region, a common optimal action exists even when intent is unknown, so there is no loss. The loss is greatest near the mean threshold, with maximum $r_{\max} = \kappa\Delta$.

Suppose that a single consultation allows the principal to communicate the correct decision for the case at hand. For now, information is not reused across cases, and the effective cost of consultation is $z \geq 0$. When only the direct cost is considered, $z = c$; when an attention constraint is introduced, the opportunity cost is included, as in Section 6.

Proposition 2 (Three decision regions). If $0 \leq z < \kappa\Delta$, the region in which consultation is strictly preferable is

$$\mathcal{Q}(z) = \left(\theta_L + \frac{z}{D(1-q)}, \quad \theta_H - \frac{z}{Dq}\right). \tag{10}$$

The AI autonomously rejects cases below this region and autonomously implements cases above it. Within the region, it follows the principal's response. The consultation rate and the minimum combined cost of consultation and residual loss are

$$\rho(z) = \left(\Delta - \frac{z}{\kappa}\right)_+, \tag{11}$$

$$R(z) = \int_0^1 \min\{r(x), z\}\, dx = \begin{cases} z\Delta - \dfrac{z^2}{2\kappa}, & 0 \le z < \kappa\Delta, \\ \dfrac{\kappa\Delta^2}{2}, & z \ge \kappa\Delta. \end{cases} \tag{12}$$

When $z \ge \kappa\Delta$, a policy without consultation is optimal. For cases at which equality holds, consultation and autonomous decision-making are equally desirable.

The proposition follows by solving $r(x) > z$ and integrating costs over the respective regions. We have $R'(z) = \rho(z)$, and R is nondecreasing and concave. As consultation becomes more expensive, it is restricted to cases in which the principal's information can avert a larger loss, so the rate at which total cost increases declines.

Even when the consultation rate is fixed at $u \in [0, \Delta]$, it is optimal to refer cases to the principal in descending order of their losses. The remaining expected loss is then

$$L(u) = \frac{\kappa}{2}(\Delta - u)^2, \qquad R(z) = \min_{0 \le u \le \Delta} \{zu + L(u)\}. \tag{13}$$

This representation allows the consultation rate and the allocation of tasks to be chosen jointly in Section 6.

5.3 A Numerical Example and Comparative Statics

Let $\theta_L = 0.3$, $\theta_H = 0.7$, $q = 0.5$, $D = 100$, and $c = 2$. These are assumed values for illustration. They imply $\mu = 0.5$, $\kappa = 25$, and $\Delta = 0.4$, so $r_{\max} = 10$ and the consultation region is $(0.34, 0.66)$.

Case evaluation	Decision-maker	Action
$x < 0.34$	AI	Autonomously rejects
$0.34 < x < 0.66$	Principal	Decides whether to implement according to the threshold
$x > 0.66$	AI	Autonomously implements

At the endpoints, the two policies are equally desirable. The consultation rate is 0.32, the consultation cost is 0.64, and the residual loss is 0.08, giving a total of 0.72. Without consultation, the expected loss is 2.

Holding the instruction and other conditions fixed, a reduction in the consultation cost expands the range of cases referred to the human. An increase in the importance parameter D also expands this range. This result is conditional on holding the instruction fixed, however. If greater importance induces a switch to complete ex ante specification, consultation at the implementation stage may decrease. Measuring human involvement therefore requires accounting for the time spent setting criteria and writing instructions, as well as the number of responses.

6 Instruction Design and the Allocation of Tasks Between Humans and AI

6.1 The Three Regimes Compared

Section 5 considered decision boundaries after delegation to AI. This section chooses who should perform the task in the first place. Evaluating and processing each case is a common requirement across regimes; after processing it, a decision is made on whether to implement or reject it. We compare the following three regimes.

Regime	Description	Direct cost and loss	Human time
H: Direct human performance	The principal uses the threshold to evaluate, decide, and process	h	τ_H
C: AI delegation with complete instructions	The threshold is communicated in advance, and the AI decides and processes	$d + k$	τ_C
I: AI delegation with incomplete instructions	The AI processes cases and consults the principal only when needed	$d + cu + L(u)$	$\tau_Q u$

Here, d is the AI processing cost incurred for each case regardless of whether it is implemented or rejected, k is the cost of fully communicating the threshold, and c is the direct cost of one consultation. The time required to prepare the instruction is τ_C, and

the time required for one response is $\tau_Q > 0$. Costs are measured in the same units of the principal's utility and do not already incorporate the scarcity value generated by a common time constraint.

In the baseline model, both the human and an AI that has been informed of the threshold make and implement the correct decision. The human processing cost is h. Human error can be accommodated by adding to h an expected loss that is independent of the regime being compared and of the allocation of attention. If fatigue or the amount of attention affects error, however, that relationship must be modeled.

The regime is selected as a rule independent of the threshold, before the principal's threshold and the case evaluation are realized. The principal then learns the threshold. Under C, the principal communicates it before the case evaluation is realized; under I, the principal responds only for cases requiring consultation. This ex ante commitment rules out the possibility that the absence of an instruction itself reveals the threshold. If the principal can freely choose the regime after learning the threshold, the analysis must also account for inferences from the regimes not chosen.

6.2 The Boundary Without an Attention Constraint

Proposition 3 (Conditions for task delegation). In the absence of an attention constraint, when the three regimes above are available, minimum cost is

$$W = \min\{h,\ d + k,\ d + R(c)\} = \min\{h,\ d + \min[k, R(c)]\}. \tag{14}$$

Thus, delegation to AI is strictly preferable if

$$d + \min\{k, R(c)\} < h. \tag{15}$$

Conditional on delegation, complete instructions are chosen if $k < R(c)$, and incomplete instructions with optimal consultation are chosen if $k > R(c)$. At equality, multiple regimes are optimal.

This result compares the minimum costs of the respective regimes. The boundary depends not only on the AI processing cost d, but also on the cost of incorporating the principal's threshold into the AI's decisions. Even if AI can process a task, the principal performs it directly when this combined cost exceeds the cost of direct performance.

In the example from Section 5, $R(c) = 0.72$. If $d = 0.10$, $k = 1.00$, and $h = 0.90$, delegation with incomplete instructions has the lowest cost, at 0.82. Reducing only k to 0.50 makes delegation with complete instructions the least costly regime, at 0.60. Starting from the initial conditions and reducing only h to 0.70 leads to direct human performance. Each of the three regimes can therefore be selected, depending on the other conditions.

6.3 A Common Attention Constraint

Human processing, instruction preparation, and responses to consultations may compete for the same person's time. Let there be finitely many task types, $j = 1, \dots, J$, with shares $w_j > 0$ satisfying $\sum_j w_j = 1$. Each type has parameters $D_j, q_j, \Delta_j, d_j, k_j, c_j, h_j$ and associated time requirements. The principal's threshold is independent across cases, and answers given for other cases do not generate learning. Persistent thresholds are analyzed separately in Section 7.

For each type, let $p_{Hj}, p_{Cj}, p_{Ij} \geq 0$ denote the shares allocated to the respective regimes, with their sum equal to one. These shares describe the allocation of a group of cases of the same type across regimes before thresholds and case evaluations are realized. Let u_j be the consultation rate among cases with incomplete instructions, and let $y_j = p_{Ij} u_j$ be the consultation rate across all cases of that type.

If the average human time available is $B \geq 0$, the design problem is

$$\begin{aligned} \min_{p,y} \quad & \sum_j w_j \left[p_{Hj} h_j + p_{Cj}(d_j + k_j) + p_{Ij} d_j + c_j y_j + \frac{\kappa_j (p_{Ij}\Delta_j - y_j)^2}{2 p_{Ij}} \right] \\ \text{s.t.} \quad & \sum_j w_j \{ p_{Hj} \tau_{Hj} + p_{Cj} \tau_{Cj} + \tau_{Qj} y_j \} \leq B, \\ & p_{Hj} + p_{Cj} + p_{Ij} = 1, \quad p_{mj} \geq 0, \\ & 0 \leq y_j \leq p_{Ij} \Delta_j. \end{aligned} \tag{16}$$

When $p_{Ij} = 0$, we have $y_j = 0$, and the fractional term is defined to be zero. This term is the residual loss in equation (13), weighted by the share of cases with incomplete instructions. The average-time constraint represents either a large population of cases or an ex ante expected time budget. It does not capture queues at particular moments or hard deadlines that must be met in every realization.

Proposition 4 (Joint allocation under an attention constraint). Suppose $B > 0$ and all costs and time requirements are finite and nonnegative. An optimum of problem (16) is characterized by the following rules for some shadow price of attention $\lambda \geq 0$:

$$\begin{aligned} G_{Hj}(\lambda) &= h_j + \lambda\tau_{Hj}, \\ G_{Cj}(\lambda) &= d_j + k_j + \lambda\tau_{Cj}, \\ G_{Ij}(\lambda) &= d_j + R_j(c_j + \lambda\tau_{Qj}). \end{aligned} \tag{17}$$

For each type, only regimes that minimize (17) are used. Cases with incomplete instructions are referred to the principal when

$$r_j(x) > c_j + \lambda\tau_{Qj} \tag{18}$$

Mixing is permitted when regimes have equal costs. An allocation satisfying the time constraint and

$$\lambda \left[B - \sum_j w_j \{ p_{Hj}\tau_{Hj} + p_{Cj}\tau_{Cj} + \tau_{Qj} y_j \} \right] = 0 \tag{19}$$

is globally optimal. Conversely, an optimal allocation can be supported by such a price and these rules.

The result follows from finite-dimensional convex optimization and its dual. Equation (17) determines who performs the task, while equation (18) allocates decisions within delegated tasks. Using the same λ in both makes the opportunity costs of direct human performance, instruction preparation, and responses consistent. When $B = 0$, regimes requiring no human time and policies without consultation are compared directly.

The shadow price is an opportunity cost used to guide choices; it is not an actual expenditure to be added again to (16). For example, setting $z = c + \lambda\tau_Q$, the actual cost of a case with incomplete instructions is $d + R(z) - \lambda\tau_Q\rho(z)$.

6.4 What Changes When Attention Becomes Scarcer?

Holding the type and regime fixed, an increase in λ weakly reduces the consultation rate. Whether it favors complete or incomplete instructions, however, depends on the human time used by each.

Proposition 5 (Substitution between instruction preparation and responses). Holding costs and time requirements fixed,

$$\frac{d}{d\lambda}\{G_C(\lambda) - G_I(\lambda)\} = \tau_C - \tau_Q \rho(c + \lambda\tau_Q). \tag{20}$$

Consequently, where complete instructions save time relative to the currently optimal consultation policy, greater scarcity of attention makes complete instructions relatively more attractive. If instruction preparation takes more time, the relationship is reversed. This is a result about the relative costs of the two regimes. The ultimate choice, which also includes direct human performance, requires comparing all of (17).

Add $d = 0.10$, $k = 1.00$, $h = 1.40$, $\tau_Q = 1$, $\tau_C = 0.05$, and $\tau_H = 0.60$ to the numerical example in Section 5. Even with a single type of case, the following allocations are optimal at different attention capacities.

Average time limit B	Optimal regime	Time used	Actual cost
$B \geq 0.32$	Incomplete instructions with optimal consultation	0.32	0.82
$B = 0.05$	Complete instructions	0.05	1.10
$B = 0$	Incomplete instructions without consultation	0	2.10

The middle row can be verified, for example, by noting that at $\lambda = 2$, $G_C = 1.20 < G_I = 1.38 < G_H = 2.60$. As capacity changes gradually, mixtures of regimes may also arise. The table does not imply that allocations across all capacities fall into precisely three stages.

In this example, as attention becomes scarce, work initially shifts from ex post responses to ex ante specification. If even the time needed to prepare instructions becomes unavailable, however, that regime can no longer be used. Thus, there is no general conclusion that instructions necessarily become more complete as human time becomes scarcer.

6.5 Lower AI Costs and Expanded Delegation

When AI processing costs fall, a more limited monotonicity result can be established.

Proposition 6 (A common AI cost and aggregate delegation). Let the AI processing cost for each type be $d_j = \bar{d}_j + \eta$, holding all other costs, information, time requirements, capacity, and feasible regimes fixed. At an optimum of (16), a decrease in the common cost η weakly increases aggregate AI delegation,

$$M = \sum_j w_j(p_{Cj} + p_{Ij}) \tag{21}$$

The result follows by comparing the optimality inequalities at two cost levels. It does not imply that a particular type must shift to AI or that consultations with humans must decrease. Increased AI delegation may raise the demand for responses, changing the instructions or the allocation of other task types. If capabilities or time requirements change along with processing costs, the assumptions of this comparative-static result no longer hold.

Cases in which AI performance changes human information-acquisition effort, as analyzed by Athey, Bryan, and Gans (2020, §III), are also outside the scope of this proposition. Proposition 6 concerns a reduction in processing costs under a fixed information structure; it does not imply that better-performing AI necessarily improves organizational outcomes.

6.6 The Meaning of the Two Boundaries

The boundary determining who performs a task is given by the comparison between G_H and $\min\{G_C, G_I\}$. Within tasks delegated to AI, the boundary determining when the human makes the decision is given by (18) under incomplete instructions. The former allocates tasks; the latter allocates decisions for individual cases.

The model therefore does not imply a fixed division of labor in which humans set objectives and AI implements the means. The principal may perform everything directly, the AI may make decisions after receiving the threshold, or the principal may respond only to exceptions. Being the source of information about the criterion is distinct from participating in every decision.

The model does not optimize permission for each external operation or contracts with AI providers. The human-AI boundary here is the boundary among the specified regimes of case processing and information transmission. It does not directly explain firm ownership, employment institutions, or the disappearance of entire occupations.

7 Repeated Delegation and Testable Implications

7.1 When Answers Are Useful for Later Cases

Section 6 assumed that a different threshold is realized for each case and that answers cannot be transferred to other cases. If the same principal continues to use the same purchasing threshold, however, a single clarification can also inform future decisions. To isolate this effect, this section removes the attention-capacity constraint and analyzes a setting in which the threshold θ remains unchanged across N cases.

Case evaluations x_t follow independent uniform distributions. A single consultation reveals the threshold exactly, and AI remembers the answer. No additional information about the threshold is obtained from the outcomes of cases in which no consultation occurred or from the principal's reactions to them. This assumption distinguishes merely observing acceptance decisions or external facts from observing the principal's utility.

The regime is chosen at the outset. Under H, the principal handles all cases. Under C, the threshold is communicated at the beginning at a one-time cost k. Under I, AI observes cases and chooses when to consult for the first time; after learning the threshold, it makes decisions autonomously. The set of regimes in this section does not include switching to direct human performance midway or purchasing partial information.

7.2 The Future Value of Consultation

Let $V_n(c)$ be the minimum expected cost under the incomplete-instruction regime when the threshold is still unknown and n cases remain. This excludes ordinary AI processing costs. With no remaining cases, $V_0(c) = 0$.

Without consultation in the current case, an expected loss $r(x)$ is incurred, and the threshold remains unknown for the next case. Consultation costs c and eliminates losses due to uncertainty about the threshold in both the current and future cases. Hence,

$$\begin{aligned} V_0(c) &= 0, \\ V_n(c) &= \int_0^1 \min\{r(x) + V_{n-1}(c), c\}\, dx \\ &= V_{n-1}(c) + R(c - V_{n-1}(c)). \end{aligned} \tag{22}$$

Proposition 7 (Reusing Answers and Ex Ante Specification). Under $c > 0$, $\kappa > 0$, and $0 < \Delta < 1$,

$$0 < V_1(c) < V_2(c) < \cdots < c, \qquad \lim_{n\to\infty} V_n(c) = c, \qquad V_n(c) \leq nR(c). \tag{23}$$

With n cases remaining, the condition for consultation is

$$r(x) > c - V_{n-1}(c) \tag{24}$$

The consultation region expands weakly as the number of remaining cases increases. At the initial regime-selection stage, it suffices to compare

$$\min\{Nh,\ Nd + k,\ Nd + V_N(c)\} \tag{25}$$

The right-hand side of equation (24) is the current consultation cost less the future cost avoided by learning the threshold. Even if an answer has little value for the current case alone, consultation can be chosen because it improves subsequent decisions. This effect does not arise when repeated delegation is treated as a series of independent one-shot problems.

Conditional on delegation to AI, complete instructions are less costly than incomplete instructions if $k < V_N(c)$. If $0 < k < c$, complete instructions become preferable for a sufficiently large number of cases. If $k \geq c$, however, complete instructions are more costly than incomplete instructions for every finite number of cases. Because a single answer can be remembered, the additional cost of incomplete instructions remains below c.

The ability to reuse the same threshold therefore does not raise only the value of ex ante specification. It also raises the value of ex post answers. The simple expression $NR(c)$ is the cost when answers are not reused, and is an upper bound on the cost of the incomplete-instruction regime in this model.

Using the same numerical values as in Section 5 gives $V_1(2) = 0.72$, $V_2(2) = 1.199232$, and $V_3(2) \simeq 1.506715$. If the cost of full specification is $k = 1$, the comparison between AI regimes selects incomplete instructions for one case and complete instructions for two cases. If $k = 2.5$, by contrast, full specification in advance does not become advantageous as the number of uses increases. In either case, comparison with direct human performance is made separately using equation (25).

This section does not jointly solve the static common-attention constraint and dynamic learning. Including daily response capacity, changing thresholds, or memory errors would require a sequential allocation problem incorporating these as state variables.

7.3 Testable Implications

The paper's predictions can be examined in experiments that distinguish the information conveyed by instructions from AI's execution capabilities. For example, after a task-design regime is chosen, the principal could be randomly assigned one of two thresholds. AI would receive an accurate case evaluation, while the conditions for communicating the threshold, consultation costs, instruction-preparation costs, and the upper limit on human time would be varied. This avoids the possibility that regime choice itself reveals the threshold.

First, experiments holding the regime fixed can test whether consultation is concentrated in cases with high $r(x)$. The predictions are that raising the stakes of a decision expands the consultation region, whereas raising the direct consultation cost contracts it. Loss should be measured to include both erroneous execution and the rejection of cases that should have been executed. Separately assessing performance when the full threshold is communicated makes it easier to distinguish missing information about intent from deficiencies in factual recognition or execution capabilities.

Second, experiments allowing free regime choice should apply the same time budget to direct human performance, instruction preparation, and responses. An experiment that limits only responses while leaving instruction-preparation time unconstrained cannot

test the substitution addressed by Propositions 4 and 5. The main outcome measures are actual costs, forgone benefits, the three types of human time, and the adoption shares of the respective regimes.

Third, experiments can distinguish a condition in which the same threshold is retained from one in which it is updated each time, and vary whether remembered answers are available. When the principal's threshold is fixed and answers can be remembered, the probability of consultation in the initial cases is predicted to increase weakly with the number of remaining cases. The number of consultations in the static setting and the timing of the first consultation in the dynamic setting should be evaluated separately.

These are conditional predictions of the model, not established empirical findings. If observations do not match the predictions, it is necessary to distinguish undetermined objectives, incorrect AI beliefs, inaccurate answers, and unobserved costs. Changing the evaluation criterion after observing outcomes would prevent identification of the loss due to missing information.

7.4 Limitations of the Model

First, the principal's objective is already determined. The analysis does not cover the process through which multiple stakeholders form objectives, the principal's discovery of their own wishes, externalities, or social legitimacy. Accurately implementing the principal's criterion is not the same as producing a socially desirable outcome.

Second, AI makes cooperative decisions under a correct prior. The model does not include unknown independent objectives, strategic concealment of information, or the incentives of AI providers. Nor does it endogenize human effort to acquire new information needed to answer a consultation. A closer connection with Athey, Bryan, and Gans (2020) would require a structure in which decision authority and consultation policies change human incentives to acquire and verify information, with consequences for answer quality and cost. This is a future extension; the current cost comparison should not be interpreted as already incorporating these incentive effects.

Third, the technologies for instruction and consultation are restricted. The communication cost k is not derived from linguistic length or cognitive processes, and the paper does not solve a general information-design problem encompassing partial instructions

or rule retrieval. If humans and AI jointly develop the processing method itself, the technology assumed to be common across regimes also changes.

Fourth, the three decision regions and the closed-form solutions depend on a binary threshold, a uniform distribution of cases, linear utility, and accurate answers. The principle that survives in more general environments is the comparison of the loss reduction from additional information with its acquisition cost and the opportunity cost of attention. The consultation set need not always be a single interval.

Fifth, the literature comparison undertaken here is intended to clarify the objects of analysis and the intellectual connections to prior work. Some original-source material has not been checked against full texts or original page images, and the full texts of some closely related studies remain unverified. Further comparison is therefore needed to assess the paper's originality. It is not presented as a finished manuscript whose relationship to existing theory has been conclusively established.

8 Conclusion

Building on the analysis of ex ante specification, information transmission, and ex post judgment in contract theory and organizational economics, this paper formulates instructions from humans to AI as a design problem. Complete instructions are defined as those for which the candidate intents consistent with the instruction and shared context imply the same permissible-action sets in the relevant situations. Clearly authorized discretion is not incompleteness. Even under incomplete instructions, a common optimal action makes it possible to avoid loss without identifying intent.

The model distinguishes two boundaries between humans and AI: who is assigned a task, and which cases within a delegated task are referred back to the principal. The former is determined by comparing the cost of direct human performance with the combined costs of AI processing, instruction, consultation, and residual loss. The latter is determined by comparing the loss reduction enabled by the principal's information with the direct cost of consultation and the opportunity cost of attention. Choosing both under the same attention constraint also reveals the substitution between instruction preparation and ex post responses.

Complete instructions are not always desirable. If preparation is costly, the chosen regime leaves instructions incomplete and refers only the necessary cases back to the

principal. Moreover, even when attention becomes scarce, a shift toward complete instructions is not guaranteed if instruction-preparation time is also scarce. When the same threshold can be used repeatedly, both ex ante specifications and ex post answers can be reused; long-term use therefore need not lead to complete instructions.

Human involvement consequently cannot be captured solely by whether each case receives approval. Performing a task directly, communicating a threshold in advance, and answering questions about selected cases are alternative ways to use the principal's information at different times and through different methods. The task of instruction theory is to specify these alternatives and their costs, and to explain the conditions under which each division of work is desirable.

Appendix A Proofs

A.1 Proposition 1

Fix the history and situation, and let the loss from action a under intent θ be $\ell(a,\theta) = V^*(s,\theta) - U(a,s,\theta) \geq 0$. If there is a common optimal action a^*, then $\ell(a^*,\theta) = 0$ for every intent with positive probability, and its expected loss is zero.

Conversely, if an action has zero expected loss, each term $p(\theta \mid h)\ell(a,\theta)$ is nonnegative and every candidate intent has positive probability, so the loss is zero for every candidate. The action is therefore a common optimal action. If the intersection is empty, each of the finitely many actions has strictly positive expected loss, and their minimum is also strictly positive. The expected loss of a randomized action is a convex combination of these losses, so randomization cannot improve on the minimum. □

A.2 Lemma 1 and Proposition 2

The cost without consultation is $r(h)$, whereas the cost with consultation is $c(h)+r^Q(h)$. Comparing the two establishes Lemma 1.

In the binary-threshold model, the expected utility of AI when it does not know the threshold is $Da(x-\mu)$, so the acceptance boundary is μ. For $\theta_L < x < \mu$, AI declines to act and forgoes the benefit $D(x-\theta_L)$ with probability $1-q$, the probability that the principal's threshold is low. For $\mu \leq x < \theta_H$, it acts and incurs the loss $D(\theta_H - x)$

with probability q, the probability that the threshold is high. Outside these regions, a common optimal action can be selected. This yields equation (9).

Solving the inequalities $r(x) > z$ on either side gives equation (10). The interval has length $\Delta - z/\kappa$, which is positive if and only if $z < \kappa\Delta$. The areas of the two end triangles left without consultation are

$$\frac{z^2}{2D(1-q)} + \frac{z^2}{2Dq} = \frac{z^2}{2\kappa}. \tag{A1}$$

Adding the consultation cost $z(\Delta - z/\kappa)$ gives the first line of equation (12). The area when there is no consultation is $\kappa\Delta^2/2$, yielding the second line.

For a fixed consultation share u, a policy that consults on a lower-loss case but not on a higher-loss case can reduce costs by exchanging the two. Consultation therefore begins with the highest-loss cases. Letting the boundary height be b, we have $u = \Delta - b/\kappa$. Replacing z with b in equation (A1) gives residual loss $\kappa(\Delta - u)^2/2$. Minimizing this loss plus consultation costs yields equation (13). □

A.3 Proposition 3

The expected cost of H is h, and that of C is $d + k$. Under I, optimal consultation gives an additional cost $R(c)$, so the total is $d + R(c)$. Taking the minimum of the three gives equations (14) and (15). Because the principal's full-information case value is common across the regimes compared, comparing costs and losses is equivalent to comparing expected utility. □

A.4 Proposition 4

The residual-loss term in equation (16) is the perspective of the convex function $u \mapsto \kappa_j(\Delta_j - u)^2/2$, weighted by p_{Ij}. On the domain $0 \le y_j \le p_{Ij}\Delta_j$, its continuous extension to $p_{Ij} = 0$ can be defined as zero. The objective is convex, the constraints are linear, and the feasible set is compact, so an optimum exists.

When $B > 0$, processing every case under incomplete instructions without consultation uses zero time. Mixing in sufficiently small shares of the other regimes and consultation then satisfies the time constraint strictly in the relative interior of the domain. Slater's condition therefore holds, and strong duality and the Karush-Kuhn-Tucker optimality conditions apply.

Let $\lambda \geq 0$ be the multiplier on the time constraint. Within cases assigned to incomplete instructions, minimization gives

$$\min_{0 \leq u_j \leq \Delta_j} \left\{ (c_j + \lambda \tau_{Qj}) u_j + \frac{\kappa_j}{2} (\Delta_j - u_j)^2 \right\} = R_j(c_j + \lambda \tau_{Qj}). \tag{A2}$$

The optimal consultation share is $\rho_j(c_j + \lambda \tau_{Qj})$. The remaining minimization over regime shares is linear, so positive shares are assigned only to regimes minimizing equation (17). The dual problem is

$$\max_{\lambda \geq 0} \left\{ \sum_j w_j \min[G_{Hj}(\lambda), G_{Cj}(\lambda), G_{Ij}(\lambda)] - \lambda B \right\}. \tag{A3}$$

Because incomplete instructions without consultation are always feasible, the sum is bounded above by $\sum_j w_j (d_j + \kappa_j \Delta_j^2 / 2)$. For $B > 0$, the objective in equation (A3) tends to $-\infty$ as $\lambda \to \infty$, and it has a finite maximizer. The proposition follows from optimality and complementary slackness.

Sufficiency can also be verified directly. An allocation satisfying equations (17) and (18) minimizes the Lagrangian for the same λ. If it is feasible and satisfies equation (19), the Lagrangian lower bound equals the actual cost of the allocation, and no feasible allocation can have a lower cost. □

A.5 Proposition 5

By $R'(z) = \rho(z)$ and the chain rule, $G_I'(\lambda) = \tau_Q \rho(c + \lambda \tau_Q)$. Since $G_C'(\lambda) = \tau_C$, subtracting yields equation (20). At $z = \kappa \Delta$, the left and right derivatives also coincide at zero. At $\lambda = 0$, the derivative is interpreted as a right derivative. □

A.6 Proposition 6

Write a complete feasible allocation as π and its cost excluding η as $F(\pi)$. For two cost levels $\eta_1 < \eta_2$, let π_1, π_2 be any respective optimal allocations. Because their feasible sets are identical,

$$\begin{aligned} F(\pi_1) + \eta_1 M(\pi_1) &\leq F(\pi_2) + \eta_1 M(\pi_2), \\ F(\pi_2) + \eta_2 M(\pi_2) &\leq F(\pi_1) + \eta_2 M(\pi_1). \end{aligned} \tag{A4}$$

Adding these inequalities gives $(\eta_2-\eta_1)[M(\pi_2)-M(\pi_1)] \leq 0$. Hence $M(\pi_1) \geq M(\pi_2)$. There is no need to hold the price of attention fixed. □

A.7 Proposition 7

If current consultation fully reveals the threshold, all subsequent losses due to uncertainty about the threshold are zero. Without consultation, no new information is obtained, and the optimal continuation cost is $V_{n-1}(c)$. Because the current action does not change future information, the minimization can be separated into the current loss and the continuation cost, yielding the second line of equation (22).

For $z > 0$, the set of cases with strictly positive loss has positive measure, so $R(z) > 0$. Conversely, the set of cases with zero loss has positive measure, so $R(z) < z$. Starting from $V_0 = 0 < c$, induction gives $z_n = c - V_{n-1} > 0$ and

$$V_{n-1} < V_n = V_{n-1} + R(z_n) < V_{n-1} + z_n = c \tag{A5}$$

In this range, $\min\{r(x)+V_{n-1}, c\} = V_{n-1}+\min\{r(x), c-V_{n-1}\}$, establishing the third line of equation (22).

Since V_n is increasing and bounded above by c, a limit $\bar{V}$ exists. Continuity of R implies $R(c-\bar{V}) = 0$, and $R(z) > 0$ for $z > 0$ implies $\bar{V} = c$. Moreover, each period's increment satisfies $R(c-V_{n-1}) \leq R(c)$, which gives $V_n \leq nR(c)$.

Equation (24) follows by comparing consultation with no consultation. Because V_{n-1} increases with n, the consultation threshold decreases. The total costs of the regimes are Nh, $Nd+k$, and $Nd+V_N(c)$, giving equation (25). The switch when $0 < k < c$ follows from $V_N \to c$, and the advantage of incomplete instructions when $k \geq c$ follows from $V_N < c$. □

Appendix B Limiting Cases and Interpretation

A known threshold. If $q \in \{0, 1\}$ or $\Delta = 0$, missing information about the threshold causes zero loss. Rather than substituting degenerate values into expressions that divide by κ or Δ, compare the alternatives directly with $R = 0$.

Costless consultation. If $c = 0$ and there is no capacity constraint, consultation on cases with divergent optimal decisions can reduce loss to zero. In the dynamic model,

$V_n(0) = 0$ as well. Even with zero direct cost, the effective cost $\lambda\tau_Q$ can be positive when attention is scarce.

Zero cost of full specification. If $k = 0$ and there is no capacity constraint, complete instructions are weakly optimal among AI regimes. If $\tau_C > 0$ and the attention constraint binds, however, their adoption cannot be inferred from monetary costs alone.

Additional information. If using a correct additional signal is optional, the minimum expected loss in the same decision problem cannot increase in ex ante expectation. Net benefits including information-acquisition costs, losses after individual signal realizations, and the allocation of human time must be assessed separately.

Regime choice after the principal learns the threshold. Choosing not to issue an instruction or consult can itself be informative. Removing the paper's ex ante commitment requires jointly determining regime choice and the posterior distributions induced by messages.

Execution failures. Complete instructions do not eliminate operational errors, misperception of facts, or failures of external systems. If these losses differ across regimes, their differences must enter the cost comparison rather than being canceled as common costs.

References

Aghion, Philippe, and Jean Tirole. 1997. "Formal and Real Authority in Organizations." *Journal of Political Economy* 105(1): 1-29. DOI.

Ante, Lennart. 2026. "Specification Cost: Relocating the Binding Constraint in Delegation to AI Agents with Externally Supplied Objectives." *Technology in Society* 88: 103503. DOI.

Athey, Susan C., Kevin A. Bryan, and Joshua S. Gans. 2020. "The Allocation of Decision Authority to Human and Artificial Intelligence." *AEA Papers and Proceedings* 110: 80-84. DOI. The text was checked using the five-page version made publicly available by an author.

Coase, R. H. 1937. "The Nature of the Firm." *Economica* 4(16): 386-405. DOI.

Dong, Jing, Prakirt Raj Jhunjhunwala, and Yash Kanoria. 2026. "Right-Sizing Communication and Recommendation Set Size in AI-Assisted Search." Working paper, arXiv:2605.23944v1. Version consulted.

Dong, Yijiang River, Tiancheng Hu, Zheng Hui, Caiqi Zhang, Ivan Vulić, Andreea Bobu, and Nigel Collier. 2026. "Value of Information: A Framework for Human-Agent Communication." *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*: 42879-42896. DOI.

Fudenberg, Drew, and Annie Liang. 2026. "Friend or Foe: Delegating to an AI whose Alignment is Unknown." Working paper, June 22. Author manuscript consulted.

Garicano, Luis. 2000. “Hierarchies and the Organization of Knowledge in Production.” *Journal of Political Economy* 108(5): 874-904. DOI.

Grossman, Sanford J., and Oliver D. Hart. 1986. “The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration.” *Journal of Political Economy* 94(4): 691-719. DOI.

Hadfield-Menell, Dylan, and Gillian K. Hadfield. 2019. “Incomplete Contracting and AI Alignment.” *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.* DOI.

Holmström, Bengt. 1979. “Moral Hazard and Observability.” *Bell Journal of Economics* 10(1): 74-91. DOI.

Maskin, Eric, and Jean Tirole. 1999. “Unforeseen Contingencies and Incomplete Contracts.” *Review of Economic Studies* 66(1): 83-114. DOI.

Simon, Herbert A. 1951. “A Formal Theory of the Employment Relationship.” *Econometrica* 19(3): 293-305. DOI.

Tirole, Jean. 1999. “Incomplete Contracts: Where Do We Stand?” *Econometrica* 67(4): 741-781. DOI.

Tirole, Jean. 2009. “Cognition and Incomplete Contracts.” *American Economic Review* 99(1): 265-294. Published-version DOI. The version cited in the text is the author's draft dated March 25, 2008; all differences from the published version have not been checked.

Note on verification of original sources. Citations have been limited to claims in the texts that could be confirmed in the reading records. Only the obtained introductory portions of Coase (1937) and Grossman and Hart (1986) were used, and no reliance is placed on page 744, which was missing from the obtained material for Tirole (1999). Equations, figures, and tables in some other original sources also remain unchecked against the original page images. The propositions in this paper do not reuse those unverified equations; they are derived in Appendix A from the assumptions explicitly stated in the text. The scope of the original-source reading and the scope of the checks on this model are recorded separately in the accompanying revision and analysis record.