NBER WORKING PAPER SERIES

AI AS STRATEGIST

Joshua S. Gans

Working Paper 33650
http://www.nber.org/papers/w33650

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2025

Thanks to Claude Sonnet 3.7 and Gemini 2.5 Pro (Experimental) for helpful research assistance. Responsibility for all errors remains my own. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

AI as Strategist
Joshua S. Gans
NBER Working Paper No. 33650
April 2025
JEL No. D81, L20, O32

ABSTRACT

This paper examines the role of artificial intelligence as a strategist in organizational decision-making by extending van den Steen's formal theory of strategy. A mathematical model is developed comparing AI and human strategists across different decision contexts, focusing on how each type generates confidence, achieves agreement, and implements decisions through control versus influence. The analysis presumes that AI excels in data-rich domains but faces credibility challenges in judgment-intensive contexts, creating a counterintuitive result where AI requires less formal authority precisely where it demonstrates superior analytical capabilities. The paper identifies distinct mechanisms through which strategic value is created: direct decision quality improvement and enhanced coordination. The authors propose domain-contingent approaches to AI integration, including differentiated authority systems across decision types and progressive control models that evolve as AI demonstrates effectiveness. These findings contribute to strategy theory while providing practical guidance for organizations seeking to effectively integrate AI into strategic processes, highlighting that organizations must adapt to their strategists' capabilities as much as strategists must match their organizations.

Joshua S. Gans
Rotman School of Management
University of Toronto
105 St. George Street
Toronto ON M5S 3E6
and NBER
joshua.gans@rotman.utoronto.ca

1 Introduction

As artificial intelligence technologies rapidly advance, questions about their potential to perform traditionally human functions have become increasingly salient. One particularly important domain is strategic decision-making. Strategy formulation has traditionally been considered a quintessentially human activity, requiring judgment, foresight, creativity, and persuasive capabilities. However, with AI systems demonstrating increasingly sophisticated capabilities in areas from pattern recognition to natural language understanding, it is reasonable to question whether AI might effectively serve as a strategist.

In exploring this, this paper must engage in some speculation regarding what an AI strategic leader might look like. This is a matter of endless debate, both academically and in popular culture, and also currently unknowable. Thus, as a starting presumption of this paper, a caricature of AI capabilities is presented. It is motivated by the following story from popular science fiction.

In the *Star Trek: The Next Generation* episode "Redemption II," Data, the android officer on the *Enterprise*, assumes command of the starship *USS Sutherland*. Initially facing skepticism from his human subordinates due to his artificial nature, Data's strategic capability is questioned, particularly when he makes critical decisions involving trade-offs to people's lives without fully articulating his reasoning. This leads a subordinate to challenge his judgment, perceiving undue risks being taken as underappreciated by the android. Only after observing the outcomes of Data's actions does the subordinate understand and appreciate Data's strategy. Data's lack of explicit communication about his strategic reasoning illustrates a significant credibility challenge that AI strategists may face. Had Data recognized the necessity or relevance of transparently explaining his plans to his subordinates, the conflict and resulting skepticism might have been mitigated or avoided entirely. This scenario emphasizes potential limitations inherent in AI-driven strategic leadership, particularly regarding the critical human dimension of establishing and maintaining trust and credibility.

The presumption of this paper is that an AI strategic leader will be more analytical and data-driven than a human counterpart who is more intuitive and willing to undertake subjective guesses. Nonetheless, here an advanced future scenario is envisioned, far surpassing today's AI capabilities primarily based on predictive analytics (Agrawal et al., 2018). In this speculative future, AI is not only capable of autonomously making strategic decisions but also adept at navigating interactions and choices involving human subordinates. Could such an AI hold advantages over human strategists? If so, what implications would this have for strategic management in organizations? Consequently, this paper is essentially a thought experiment: if an artificial intelligence (AI) truly assumed responsibility for organizational strategy, what potential advantages might arise?

This paper is grounded, however, in existing understanding of strategic leadership. It formally analyzes the conditions under which AI can perform the strategist's function; grounding our analysis in a rigorous formal theory of strategy developed by van den Steen (2017). This theory provides precise definitions and characterizations of what strategy is, what makes decisions strategic, and what the strategist's essential functions are. By using this foundation, we can evaluate AI's potential as a strategist with conceptual clarity and precision.

van den Steen defines strategy as "the smallest set of choices to optimally guide other choices" (van den Steen, 2017, p. 2616). This functional definition focuses on what strategy does rather than what it contains, making it particularly suitable for analyzing whether AI can fulfill the strategist's role. In van den Steen's framework, a strategist investigates potential choices, selects which ones to announce as strategy, and thereby guides other decisions within the organization. The effectiveness of strategy depends on its reliability, the centrality of the chosen decisions, and the strength of interactions among decisions.

To address the question of whether AI can serve as a strategist, we extend van den Steen's model to explicitly incorporate an AI decision-maker alongside human participants. We model AI as having potentially different capabilities in terms of information processing, credibility, and control over decisions. This approach allows us to formally characterize the conditions under which AI can effectively perform the strategist's functions.

Our analysis yields several intriguing insights. First, we characterize how AI and human strategists systematically differ in their capabilities and limitations in line with the starting presumption that AIs have advantages in structured, data-rich environments, while human strategists maintain advantages in novel or ambiguous contexts requiring creative interpretation. Second, we formally analyze the value of strategic interventions, demonstrating that strategic value is created through two distinct mechanisms: a direct decision quality effect and a coordination effect. Importantly, the relative core capabilities of AI and human strategists may manifest themselves differently in these two effects. For example, AIs may actually have coordination advantages if their actions are sufficiently transparent. We further identify that the incremental value of control is greatest precisely when influence is least effective—when agreement is low and when there are significant differences in confidence levels between strategists and participants.

Third, we explore competitive interactions, finding that AI strategists have comparative advantages in environments with established competitive patterns and rich historical data, particularly in quantity-based competitive moves where commitment credibility is valuable. Finally, we identify a fundamental relationship between a strategist's credibility and their need for formal control—as credibility increases, the value of formal control diminishes. This creates a counterintuitive implication: in data-rich domains where AI demonstrates superior analytical capabilities, formal control becomes less necessary as agreement may naturally emerge. Our analysis suggests organizations should develop domain-contingent approaches to AI integration, with differentiated authority systems across decision types and progressive authority models that evolve as AI demonstrates effectiveness. These insights extend strategy theory while providing practical guidance for organizations considering AI integration into their strategic processes but critically highlights the notion, underexplored in the current formal literature, that not only must strategists match their organization but organizations need to match their strategists.

The paper contributes to both the strategy literature and the growing literature on AI capabilities and limitations. For strategy scholars, we provide a formal characterization of when and how AI can contribute to strategy formulation, extending existing theories of strategy to incorporate technological agents. For AI researchers and practitioners, we identify the specific capabilities that AI systems would need to develop to effectively serve as strategists, highlighting areas where current technologies fall short.

These results are particularly interesting because they shift the debate from a simplistic question of whether AI can replace humans to a deeper exploration of the core components of strategic effectiveness. The paper emphasizes that strategic success is contingent not merely on accurate analysis, but also fundamentally on the strategist's capacity to generate credibility and secure organizational commitment.

Consequently, the insights presented here not only inform how organizations might better leverage AI technologies but also significantly enrich our understanding of the essential qualities required for effective strategic leadership.

1.1 Literature Review

van den Steen (2017) develops a formal theory of strategy that seeks to provide rigorous answers to fundamental questions about what strategy is and why it matters. Unlike many existing approaches in the strategy literature that offer descriptive definitions based on what strategy looks like, van den Steen proposes a functional definition based on what strategy does: "the smallest set of choices to optimally guide (or force) other choices" (van den Steen, 2017, p. 2616).

This definition captures the core insight that strategy's purpose is to ensure that an organization's choices fit together coherently, both at a point in time and over time. A strategy is thus like a plan reduced to its essence—the minimum set of choices needed to guide other decisions in the organization. The definition implies that strategy is not about specifying every detail but about providing just enough guidance to ensure coherence across decisions.

In van den Steen's model, a project involves multiple decisions that collectively determine its outcome. Each decision-maker has local information about their own decision but limited knowledge about other decisions. In this context, a strategist can investigate and announce some choices, which then guide other decisions within the organization. The equilibrium announcement in this model coincides with the definition of strategy as the smallest set of choices to optimally guide other choices.

Based on this framework, van den Steen identifies several characteristics that make decisions strategic:

1. **Ex ante uncertainty with clear implications**: For a decision to be strategic, it should be uncertain (as obvious choices provide no guidance) but have clear implications for other decisions.

2. **Reliability**: A decision is strategic only if it is reliable, meaning that the actual choice will match what was announced in the strategy. Otherwise, other decisions will not be guided effectively.

3. **Strong interactions**: Decisions with strong interactions with other choices are more strategic because aligning on these decisions generates more value.

The theory also identifies conditions under which strategy is most valuable, including when there are many strong interactions among decisions, when decisions are irreversible, and when there is significant uncertainty or ambiguity.

In a subsequent paper, van den Steen (2018a) extends his formal theory to analyze how the identity and characteristics of the strategist affect strategy formation and execution. This work is particularly relevant for our analysis of AI as a potential strategist. van den Steen (2018a) shows that strategy formulation by the CEO or key decision-makers leads to both better strategy and better execution compared to strategy formulation by outsiders or other insiders. The key insight is that when the strategist controls the strategic decisions, this provides credibility to the strategy announcement, making it more likely that other decisions will align with it. In contrast, outsiders or insiders without control over strategic decisions face a credibility problem: others doubt that the announced strategy will be followed, reducing its effectiveness as a guide.

A particularly important element in van den Steen's analysis is the role of disagreement. The model allows for differing priors, meaning that rational individuals can openly disagree about the optimal choice even with the same information. In the presence of disagreement, the strategist's control over strategic decisions becomes critical for strategy execution. Without such control, the strategy lacks credibility, and other decisions are less likely to align with it.

While there is a growing literature on AI as a decision-maker, relatively little research has directly addressed AI's potential role in strategy formulation. Most existing work focuses on AI's capabilities for prediction (Agrawal et al., 2018), operational decision-making (Brynjolfsson and McAfee, 2017), or specific domains such as finance or healthcare (Jarrahi, 2018). Agrawal et al. (2018) characterize AI as primarily a prediction technology, reducing the cost of prediction in various domains. They argue that as prediction becomes cheaper, the value of complementary human judgment increases. This framework suggests a potential division of labor between AI and humans in decision-making contexts, but does not specifically address strategy formation.[1]

[1] Agrawal et al. (2024) actually use van den Steen's environment as a foundation for analyzing AI adoption. However, their interest is on the system-wide changes to support adoption rather than a

Some authors have discussed AI's limitations for complex decision-making involving fundamental uncertainty. Chalmers et al. (2021) argues that AI systems excel at pattern recognition in data-rich domains but struggle with novel situations requiring conceptual innovation. Similarly, von Krogh (2018) suggests that human judgment remains essential for decisions involving values, ethics, and tacit knowledge.

A few scholars have begun to explore AI's potential role in strategic decision-making. Shrestha et al. (2019) suggest that AI could augment human strategists by providing data-driven insights, but argue that human judgment remains essential for synthesizing these insights into coherent strategies. Raisch and Krakowski (2021) propose a framework for human-AI interaction in decision-making, identifying different modes of collaboration depending on the nature of the task.

However, none of these works provides a formal analysis of the conditions under which AI could effectively serve as a strategist. Our paper addresses this gap by extending van den Steen's formal theory of strategy to incorporate AI as a potential strategist, allowing us to precisely characterize when and how AI can perform this function.

1.2 Research Questions and Approach

Our central research question is: what consequences follow when AI can effectively perform the strategist's role? This allows us to address when an AI strategist be preferred to a human strategist and how the characteristics of decisions (data-richness, uncertainty, centrality) affect AI's effectiveness as a strategist? To address these questions, we extend van den Steen's model to explicitly incorporate AI as a potential strategist alongside human participants. We model AI as having potentially different capabilities in terms of information processing, credibility, and control over decisions.

Our approach involves developing a formal model that captures the essential features of van den Steen's framework while adding parameters specific to AI's capabilities and limitations. As such, we will be able to characterise, at the decision level, the relative effectiveness of an AI strategist.

The remainder of the paper is organized as follows. Section 2 presents our formal model of AI as strategist with some baseline results. Section 3 then compares AI versus human strategists. Sections 4 and 5 then provides a deeper exploration of

strategic perspective.

credibility and commitment by AIs, including the role of competitive interactions. A final section concludes.

2 Model Setup

We develop a formal model of strategic decision-making that explicitly captures the interplay between a strategist (human or AI) and operational managers in organizations. Our approach builds on the framework pioneered by van den Steen (2017, 2018a), which conceptualizes strategy as a mechanism for guiding organizational choices under uncertainty. This framework is particularly valuable for our purpose of comparing how different types of strategists (human versus AI) shape organizational outcomes when decisions involve subjective judgment.

As Knight (1921) observed, strategic decisions often involve "situations which are far too unique [...] for any sort of statistical tabulation to have any value for guidance." In such contexts, pure data analysis is insufficient, and rational actors may hold differing beliefs about optimal choices even with access to identical information—what economists call "differing priors" (Morris, 1995; van den Steen, 2010b).[2] This feature is central to understanding the comparative advantages of human versus AI strategists, as we will demonstrate.

2.1 Project Structure and Decision Environment

Consider an organization undertaking a project composed of K interdependent decisions, denoted $D_1, \ldots, D_K$. For each decision D_k, a specific choice d_k must be selected from a set of possibilities $\mathcal{D}_k$. The project's overall success, measured by revenue R, depends on both the individual correctness of each choice and their collective coherence. Following van den Steen (2017), we assume the choice sets $\mathcal{D}_k$ represent a continuum (or are sufficiently large) to simplify the analysis of belief alignment.

2.1.1 Stand-Alone Correctness: External Fit

For each decision D_k, we posit the existence of an unknown true state $T_k \in \mathcal{D}_k$ representing the objectively "correct" choice when that decision is considered in isolation.

[2] A related literature that explores similar issues is the literature on theory-based decision making in strategy. See, for example, Ehrig and Schmidt (2022); Felin and Zenger (2017); Chalmers et al. (2021); Felin et al. (2024); Wuebker et al. (2023).

This captures the concept of external fit—how well each decision aligns with the demands of the external environment. For example, T_k might represent the optimal product features given current market conditions, or the best production technology given cost structures.

The stand-alone contribution of decision D_k to revenue is:

$$\alpha_k \mathbf{1}(d_k = T_k) \tag{1}$$

where $\alpha_k > 0$ is the economic importance of getting this particular decision right, and $\mathbf{1}(\cdot)$ is the indicator function that equals 1 when $d_k = T_k$ and 0 otherwise. This formulation captures the idea that deviating from the optimal choice for the external environment entails an opportunity cost of α_k.

2.1.2 Internal Alignment: Coordination Requirements

While each decision has an externally optimal choice, organizational effectiveness also requires internal coherence among decisions. For example, the ideal marketing approach must align with the chosen production technology, and the talent strategy must support the selected business model.

Following van den Steen (2017), we formalize this through the concept of interaction states T_{kl}. For each pair of decisions (D_k, D_l), the interaction state T_{kl} defines the alignment requirement as a bijection (one-to-one correspondence) between the choice sets $\mathcal{D}_k$ and $\mathcal{D}_l$. Specifically, a pair of choices (d_k, d_l) is aligned if and only if $(d_k, d_l) \in T_{kl}$.

The bijection property captures an important feature of organizational dependencies: for any choice d_l made for decision D_l, there exists exactly one choice d_k for decision D_k that optimally complements it. For instance, if d_l represents a luxury market positioning strategy, d_k might need to be a high-quality, small-batch production approach to achieve alignment.

The revenue contribution from this alignment is:

$$\gamma_{kl} \mathbf{1}((d_k, d_l) \in T_{kl}) \tag{2}$$

where $\gamma_{kl} \geq 0$ represents the economic importance of achieving alignment between these specific decisions.

2.1.3 Total Revenue Function

Combining these components, the total revenue contribution associated with decision D_k is:

$$R_k = \alpha_k \mathbf{1}(d_k = T_k) + \sum_{l \in K \setminus \{k\}} \gamma_{kl} \mathbf{1}((d_k, d_l) \in T_{kl}) \tag{3}$$

The project's total revenue is the sum across all decisions:

$$R = \sum_{k=1}^{K} R_k \tag{4}$$

This payoff structure creates a fundamental strategic tension: the stand-alone optimal choice T_k may differ from the choice required to achieve alignment with other decisions, forcing trade-offs between external fit and internal coherence. This tension is what makes strategic guidance valuable.

2.2 Decision-Makers and Their Beliefs

The decisions D_k are made by participants (operational managers) P_k. A strategist S—either human (H) or AI (A)—oversees the project. The true states T_k and interaction requirements T_{kl} are initially unknown to all players, necessitating judgment under uncertainty.

2.2.1 Belief Formation About Stand-Alone Optimality

Each decision-maker forms subjective beliefs about the stand-alone optimal choice for each decision:

- Participant P_k forms a belief characterized by the pair (θ_k^P, ν_k^P), where $\theta_k^P \in \mathcal{D}_k$ represents their best estimate of T_k, and $\nu_k^P \in (0, 1)$ represents their subjective confidence that $\theta_k^P = T_k$.
- The strategist S, if they investigate decision D_k, forms a belief (θ_k^S, ν_k^S) reflecting their judgment about the optimal choice and their confidence in that judgment.

Critically, we allow for differing priors: θ_k^P may differ from θ_k^S even without any private information, reflecting different mental models, expertise, or interpretive frameworks. As van den Steen (2010a) notes, such disagreement is common in strategic contexts where the "right" approach cannot be determined purely from data. However, the paper does not model specifically the data generating process for those beliefs and hence, we are making the assumption here that the beliefs are correlated with with underlying true state , T_k, to some degree.

This approach allows us to analyze how potential disagreement between the strategist (human or AI) and operational managers affects organizational outcomes—a central concern when implementing AI-driven strategic guidance.

2.2.2 Agreement Parameter

While beliefs may differ, they will generally show some correlation. We capture this through an agreement parameter $\rho_k \in [0, 1]$, which represents the probability that participant P_k's judgment about the optimal stand-alone choice coincides with the strategist's judgment (i.e., $\theta_k^P = \theta_k^S$, conditional on S forming a belief about D_k).

This parameter likely has distinct interpretations depending on strategist type and so will play an important role in the analysis that follows.

- For a human strategist, ρ_k reflects shared mental models, organizational culture, and interpersonal rapport.
- For an AI strategist, ρ_k reflects how well the AI's recommendations align with human intuition and domain expertise.

A higher ρ_k implies more frequent agreement on the right course of action, which (as we will show) facilitates strategy implementation.

2.2.3 Knowledge of Alignment Requirements and Agreement

Following van den Steen (2017), we assume participants P_k have perfect knowledge of the interaction states T_{kl} and T_{lk} that define how their decision must align with others. This assumption reflects the practical reality that operational managers typically understand the technical dependencies between decisions (e.g., how marketing must align with production) even when they may disagree about what choices are optimal in isolation.

Central to the strategist's ability to guide decisions without formal control is the concept of agreement, captured by the parameter $\rho_k \in [0, 1]$. In the context of the influence mechanism described later (Section 2.3), ρ_k represents the probability that participant P_k will follow the strategist's announced recommendation θ_k^S when the strategist lacks formal control ($\lambda_k = 0$) and makes an announcement. This parameter reflects factors influencing the reception of strategic guidance, such as shared mental models, the perceived credibility or transparency of the strategist (human or AI), and alignment with participant intuition or expertise (Griffith, 1999; Hambrick, 2007). A higher ρ_k indicates a greater likelihood that strategic announcements will translate into aligned action, even without direct authority.[3]

The strategic challenge, therefore, involves not only navigating uncertainty about the stand-alone optimal choices T_k and achieving alignment according to the known rules T_{kl}, but also managing the effectiveness of strategic guidance through achieving sufficient agreement ρ_k.

2.3 Control, Influence and Strategy Implementation

The strategist S can affect decisions through two mechanisms: formal authority (control) or informal influence:

- **Control ($\lambda_k = 1$):** When the strategist has formal authority over decision D_k, they directly implement their preferred choice: $d_k = \theta_k^S$. Control represents decisional authority—the strategist can mandate implementation of their vision.

- **Influence ($\lambda_k = 0$):** Without formal authority, the strategist can only announce their belief θ_k^S as a form of strategic guidance (cheap talk). The participant P_k retains decision rights. In response to the strategist's announcement, the participant's action is determined as follows:

  - With probability ρ_k (the agreement parameter, reflecting the likelihood P_k follows the recommendation, as discussed in Section 2.2.3), the participant's decision aligns with the strategist's announced belief: $d_k = \theta_k^S$. This occurs because the announcement, combined with factors like strategist credibility, shared mental models, or the perceived strength of the

[3] While ρ_k represents the likelihood of following a recommendation, it is distinct from, though likely correlated with, the participants' and strategist's confidence levels (ν_k^P, ν_k^S) in their own beliefs matching the true state T_k. We assume beliefs (θ_k) are implicitly correlated with the underlying state T_k. Pathological cases, such as both agents having high confidence ($\nu > 1/2$) but zero agreement ($\rho_k = 0$), are implicitly excluded as they challenge the premise of shared context or effective communication.

strategist's reasoning (especially for AI), effectively persuades or guides the participant.

  - With probability $(1 - \rho_k)$, the participant disregards the strategist's guidance and implements their own judgment: $d_k = \theta_k^P$.

If the strategist makes no announcement regarding D_k, then P_k defaults to implementing their own belief: $d_k = \theta_k^P$.

This formulation explicitly links the effectiveness of influence to the agreement parameter ρ_k: when ρ_k is high (indicating higher credibility or better alignment of perspectives), announced strategies are more likely to be followed even without formal authority. While simpler than a full game-theoretic equilibrium, this behavioral rule captures the empirical reality that the effectiveness of influence hinges on factors affecting the reception and acceptance of strategic guidance, such as perceived judgment alignment, trust, and communication clarity (Griffith, 1999; Hambrick, 2007).

2.4 Game Structure and Strategic Process

The strategic process unfolds in three main phases, building on van den Steen (2018a)'s framework:

1. **Strategy Formulation**

   (1a) **Investigation:** The strategist S may investigate one decision state $T_{\tilde{k}}$ at cost c_S, forming a belief $(\theta_{\tilde{k}}^S, \nu_{\tilde{k}}^S)$. This step represents strategic analysis and the development of a perspective on a key issue.

   (1b) **Announcement:** The strategist may make a public announcement M (the "strategy") regarding their preferred choice. This represents the communication of strategic direction to the organization.

2. **Strategy Implementation**

   (2a) **Belief Formation:** Each participant P_k forms a belief (θ_k^P, ν_k^P) about their decision, observes the strategist's announcement M (if any), and learns the relevant interaction rules T_{kl}. Each P_k implicitly determines whether they agree with the strategist's announced view (which occurs with probability ρ_k).

(2b) **Decision Making:** Decisions d_k are made simultaneously (or without observability between participants) according to the control and influence mechanisms described earlier. This represents the implementation of strategic decisions throughout the organization.

3. **Outcomes**

(3) **Payoff Realization:** Revenue R is determined based on the choices made d_k, the true states T_k, and the alignment rules T_{kl}.

The participants aim to maximize their expected decision-specific payoff $E[R_k]$ (subject to the choice rules described above), while the strategist aims to maximize the expected total revenue minus investigation costs: $E[R] - c_S$. Please refer to Table 1 for a full list of exogenous and endogenous variables used throughout this paper.

2.5 Simplified Three-Decision Model

For analytical tractability, we focus on a setting with $K = 3$ decisions. This provides sufficient complexity to capture the key strategic tensions while remaining manageable. We further simplify by assuming uniform interaction importance $\gamma_{kl} = \gamma$ for all $k \neq l$. The parameter γ thus represents the general importance of achieving alignment across decisions relative to the stand-alone correctness values α_k.

From the perspective of overall project success, which the strategist aims to maximize, the total expected revenue $E[R]$ depends on the choices made by participants aiming to maximize their own expected payoffs $E[R_k]$ (subject to influence or control). In the simplified specification, this overall expected revenue can be reduced to:

$$R = \sum_{k=1}^{3} \alpha_k \mathbf{1}(d_k = T_k) + \gamma \sum_{1 \leq k < l \leq 3} \mathbf{1}((d_k, d_l) \in T_{kl}) \tag{5}$$

The first term captures the value of external fit for each decision, while the second term sums the alignment benefits across all pairs of decisions. By using the range $1 \leq k < l \leq 3$, we consider each decision pair exactly once, avoiding double-counting while still capturing all relevant interactions.

Despite these simplifications, the model preserves the essential features needed to compare human and AI strategists: belief formation about unknown optimal choices

T_k, known alignment requirements T_{kl}, varying degrees of agreement ρ_k, differential control λ_k, and the mechanisms linking these parameters to organizational choices d_k.

2.6 Evaluating Performance: Correctness and Alignment

To analyze the effectiveness of different strategic approaches, we examine two key outcome metrics: the probability of making correct stand-alone decisions and the probability of achieving alignment between decisions.

2.6.1 Stand-alone Correctness Probability

For each decision D_k, the probability of making the correct stand-alone choice ($d_k = T_k$) depends on whose belief determines the actual choice, and the confidence associated with that belief:

- When the strategist controls the decision ($\lambda_k = 1$):

  $$\Pr(d_k = T_k) = \nu_k^S \tag{6}$$

  This reflects that the strategist implements their judgment θ_k^S, which matches the true optimal choice T_k with probability ν_k^S.

- When the participant controls the decision ($\lambda_k = 0$) and there is no strategic announcement:

  $$\Pr(d_k = T_k) = \nu_k^P \tag{7}$$

  The participant implements their judgment θ_k^P, which is correct with probability ν_k^P.

- When the participant controls the decision ($\lambda_k = 0$) but the strategist makes an announcement:

  $$\Pr(d_k = T_k) = \rho_k \nu_k^S + (1 - \rho_k)\nu_k^P \tag{8}$$

  With probability ρ_k, the participant adopts the strategist's view (correct with probability ν_k^S); with probability $(1-\rho_k)$, they maintain their own view (correct with probability ν_k^P).

2.6.2 Alignment Decision Rules

Following van den Steen's model, we simplify the alignment mechanism by treating it as a deterministic choice rather than a probabilistic outcome. Specifically, for each decision D_k, the participant decides whether to optimize for stand-alone correctness or for alignment with another decision D_l.

For a participant P_k deciding whether to align with decision D_l:

- Choose stand-alone optimal ($d_k = \theta_k^P$) if $\alpha_k \nu_k^P > \gamma_{kl}$
- Choose to align with d_l (following T_{kl}) if $\alpha_k \nu_k^P < \gamma_{kl}$

A key simplification arises from the simultaneous nature of decisions, meaning participant P_k cannot observe d_l when choosing d_k. We follow van den Steen (2017) in simplifying this coordination aspect. The decision rule implies that if γ_{kl} is sufficiently high relative to $\alpha_k \nu_k^P$, P_k prioritizes alignment according to the known rule T_{kl}. This implicitly assumes participants can successfully coordinate (e.g., based on expectations of each other's likely actions or via unmodeled communication) when alignment is mutually perceived as beneficial, allowing us to focus on the strategic choice *between* stand-alone optimization and alignment, rather than the mechanics of coordination itself.

When the strategist announces a strategy for decision D_l, participant P_k faces an additional complexity: with probability ρ_k, their decision aligns with the strategist's recommendation (as per Section 2.3), but with probability $(1 - \rho_k)$, they make their own choice following the above stand-alone vs. alignment rule.

2.6.3 Alignment Outcomes

Since alignment is deterministic once choices are made, the probability of achieving alignment between decisions D_k and D_l depends solely on the decisions made by the relevant participants and whether those decisions satisfy the alignment requirement T_{kl}.

When both decisions are controlled by the strategist ($\lambda_k = \lambda_l = 1$), alignment is achieved with certainty if the strategist chooses to align these decisions. When one decision is controlled by the strategist and one by a participant, alignment depends on whether the participant chooses to align with the strategist's decision or to optimize for stand-alone correctness. When both decisions are controlled by participants, alignment occurs when either:

- Both participants naturally agree with the strategist's announced recommendation, or
- One or both participants explicitly choose to align with the other's decision rather than optimizing for stand-alone correctness

This simplified approach to alignment maintains the key strategic tension between external fit and internal coherence while eliminating the probabilistic alignment parameters from the original formulation.

2.7 Expected Revenue and Strategic Analysis

The expected total revenue under any strategic approach must account for the fundamental trade-off each decision-maker faces: optimizing for stand-alone correctness or for alignment with other decisions. For each decision D_k, participates aim to maximise their component of revenue while the strategist considers which objective to pursue based on the relative expected payoffs as in equation (5): $\mathbb{E}[R] = \sum_{k=1}^{3} \left[S_k \cdot \alpha_k \nu_k + \sum_{l \neq k} A_{kl} \cdot \gamma \right]$. Importantly, this formulation explicitly captures the fundamental trade-off each participant faces: for any decision k, they must either pursue stand-alone correctness ($S_k = 1$) and earn an expected payoff of $\alpha_k \nu_k$, or deliberately align with another decision ($A_{kl} = 1$ for some l) and earn a payoff of γ. The mutual exclusivity constraint ($S_k + \sum_{l \neq k} A_{kl} \leq 1$) formalizes that a participant cannot simultaneously optimize for both objectives.

This formulation allows us to compare different strategic approaches by examining:

- How strategists influence which decisions prioritize stand-alone correctness versus alignment
- How effectively strategists steer participants toward the optimal trade-off patterns across decisions
- The resulting organizational revenue, which directly reflects the strategist's ability to guide these interdependent choices toward coherent, value-maximizing outcomes

In subsequent sections, we will use this framework to analyze how human and AI strategists differ in their effectiveness across various organizational contexts, control structures, and alignment scenarios.

3 Comparing AI and Human Strategists

Having established our formal model, we now explore the key dimensions along which AI (S = AI) and human ($S = H$) strategists systematically differ. These differences—in judgment formation, agreement patterns, and decision rule characteristics—directly influence their strategic effectiveness across organizational contexts. Understanding these differences is crucial for determining when AI might enhance or potentially detract from strategic decision-making.

Our model highlights several key parameters that may differ systematically between human and AI strategists:

- **Investigation cost (c_S):** The relative efficiency of humans versus AI in conducting strategic analysis
- **Belief accuracy (ν^S):** Whether human or AI strategists form more accurate judgments about stand-alone optimal choices
- **Agreement with operational managers (ρ):** Whether human or AI strategists achieve higher agreement with those implementing decisions
- **Control effectiveness:** The relative ability of human versus AI strategists to exercise effective control over decisions

Of these, the investigation cost is something that an AI may naturally excel at but also, to the extent that they do, there is no reason why a human could not use AI in that particular function. The remaining factors are, however, less clear in the application. Hence, we will focus on those in what follows.

That said as anticipated in the Introduction, we will, in general, presume that AIs have superior analytical capabilities in dealing with data-rich decision environments while humans have advantages in terms of judgment in environments where subjective evaluations and intuition are more important. It is important to emphasise that there is a sense in which this dichotomy lacks foundation and is extrapolating from known stereotypes between machines and people. Thus, this analysis should be seen as presuming such dichotomies but also highlighting the importance of those assumptions for conclusions drawn here.

3.1 Differential Agreement Patterns

At the core of a strategist's job is bring other participants along with their plans with the minimal amount of effort; that is, their role in coordination. At the heart of this are whether agreement amongst participants can be achieved or not.[4]

The agreement parameter ρ_k^S captures the likelihood that participant P_k independently shares strategist S's belief about the optimal stand-alone choice. When a participant and strategist share similar mental models, they are more likely to agree on the appropriate course of action even without extensive communication. This agreement significantly affects strategy implementation, particularly when the strategist lacks formal authority and must rely on influence.

The starting point for considering humans versus AIs in this regard are whether it is likely that $\rho_k^H > \rho_k^{\text{AI}}$ or not. To this end, a reasonable analysis might be as follows. Human strategists typically benefit from higher baseline agreement with operational managers due to shared cognitive frameworks, organizational culture, and contextual understanding (Kahneman, 2011; Tichy and Sherman, 1993). Humans can leverage rich communication channels—including metaphors, stories, and appeals to shared experiences—to build consensus around strategic recommendations. This communication richness facilitates implementation via influence even when formal authority is limited.

By contrast, AI systems may face lower agreement in judgment-rich domains characterized by ambiguity and tacit knowledge. This stems from AI's potentially opaque reasoning processes and lack of shared experiential context with human decision-makers (Brynjolfsson and Rock, 2022). However, in data-rich, verifiable domains where recommendations can be backed by clear evidence and analysis, AI might achieve high agreement based on demonstrable performance advantages. The agreement parameter ρ_k^{AI} will likely be higher for decisions where AI can effectively communicate its reasoning through data visualization and transparent analysis.

[4] For a recent discussion of the role of disagreement in strategic buy-ins in organisations see Gans (2024).

The agreement differential between human and AI strategists has important implications: where $\rho_k^H > \rho_k^{\mathrm{AI}}$, humans may be more effective when relying on influence rather than authority, while AI might require more formal control to achieve comparable strategic impact. As will be seen in what follows (specifically in Section 5 below, this implication does not necessarily follow from a simple comparison of these parameters.

3.2 Confidence Development Patterns

The confidence parameter ν_k^S reflects the strategist's belief in the correctness of their stand-alone recommendation θ_k^S. This parameter captures both actual accuracy and the strategist's assessment of their own judgment quality. Different types of strategists develop confidence through fundamentally different mechanisms.

Another starting point for considering humans versus AIs in this regard are whether it is likely that $\nu_k^H > \nu_k^{\mathrm{AI}}$ or not. Again, a reasonable analysis might be as follows. Human confidence blends experience-based pattern recognition, tacit domain knowledge, and analytical reasoning. Humans often excel in ambiguous situations by leveraging intuition developed through years of experience (Klein, 2007). However, human judgment is susceptible to well-documented cognitive biases, including overconfidence, confirmation bias, and availability heuristics (Lovallo and Kahneman, 2003; Keating, 2012). Human confidence may be disproportionately influenced by recent experiences or emotionally salient outcomes rather than objective probabilities.

By contrast, AI confidence is typically grounded in statistical patterns identified in training data or simulations. AI excels in domains with rich, representative historical data and well-defined success metrics (Amodei et al., 2016). However, AI systems may struggle with novel situations that differ substantially from their training examples. Importantly, $\nu_k^{S,\mathrm{AI}}$ likely scales more systematically with data quantity and quality than does $\nu_k^{S,H}$, potentially enabling superior performance in data-rich environments while underperforming humans in data-scarce or highly novel contexts.

These confidence patterns suggest domain-specific advantages: AI strategists may develop more reliable confidence in structured, data-rich environments, while human strategists maintain advantages in novel or ambiguous contexts requiring creative interpretation of limited evidence.

3.3 Decision Rule and Alignment Characteristics

Rather than probabilistic alignment tendencies, it is also possible to consider how different strategist types influence the critical decision rule that determines when participants choose alignment over stand-alone optimality:

- **Threshold Clarity:** For a participant to choose alignment ($\alpha_k \nu_k^P < \gamma_{kl}$), they must clearly understand both the value of stand-alone optimality ($\alpha_k \nu_k^P$) and the value of alignment (γ_{kl}). Human strategists may excel at communicating and reinforcing these values through organizational culture and shared narratives. AI strategists might provide more precise quantification of these parameters but could struggle to establish shared understanding of their relative importance.

- **Alignment Consistency:** When making choices to align with other decisions, consistency in following the known alignment rules T_{kl} is essential. AI strategists likely demonstrate higher consistency in identifying and implementing precise alignment relationships, while human strategists might rely more on approximation and intuition, potentially leading to occasional misalignments.

- **Adaptation to Alignment Changes:** As business conditions evolve, the optimal alignment relationships may shift. Human strategists might more readily identify and adapt to fundamental changes in these relationships, particularly when they require reconceptualizing the business model. AI strategists excel at detecting incremental shifts in established alignment patterns but may struggle to recognize or implement transformative realignments.

These alignment characteristics significantly influence each strategist type's ability to improve organizational coordination outcomes. An AI strategist with high consistency in applying alignment rules but challenges in communicating threshold values would excel when given broad authority but struggle when required to influence without control.

4 The Value of Strategic Interventions

Having outlined the broad potential differences between AI and human strategists in terms of exogenous parameters, we now turn to consider how these translate and manifest themselves in consider the value of each in engaging in strategic interventions. By comparing the effectiveness of different strategic approaches (control versus influence) across various decision contexts, we derive more precise conditions under which AI versus human strategists generate superior outcomes.

4.1 Baseline: Decision-Making Without Strategic Guidance

We first establish a baseline scenario that serves as our reference point. Without strategic intervention, each participant P_j makes decisions based solely on their own judgment about the relative value of stand-alone optimality versus alignment.

Definition 1 (Baseline Decision Rule)**.** *In the baseline scenario, participant P_j makes decision d_j according to:*

- *Choose $d_j = \theta_j^P$ (stand-alone optimal) if $\alpha_j \nu_j^P \geq \gamma_{jl}$ for all $l \neq j$*
- *Choose $d_j = a_j(d_l)$ (align with decision l) if $\alpha_j \nu_j^P < \gamma_{jl}$ for some $l \neq j$, where $a_j(d_l)$ is the choice that aligns with d_l according to T_{jl}*

The baseline decision rule captures the fundamental trade-off each participant faces: whether to prioritize external fit (stand-alone optimality) or internal coherence (alignment with other decisions). As noted previously (Section 2.6.2), implementing the alignment choice ($d_j = a_j(d_l)$) presents a coordination challenge due to the simultaneous nature of decisions. Our baseline analysis, following van den Steen (2017), assumes that if participants independently determine alignment to be optimal based on the $\alpha_j \nu_j^P < \gamma_{jl}$ condition, they can successfully coordinate on an aligned outcome. This simplification allows the focus to remain on the individual trade-off decision.

The expected revenue in this baseline scenario is:

$$\mathbb{E}[R|\text{baseline}] = \sum_{j=1}^{3} \left[S_j^P \cdot \alpha_j \nu_j^P \right] + \gamma \cdot N_{\text{align}}^{\text{base}} \tag{9}$$

Where:

- S_j^P is an indicator that equals 1 if participant P_j chooses stand-alone optimization (based on Definition 1) and 0 otherwise.

- $N_{\text{align}}^{\text{base}}$ is the number of decision pairs (j, l) for which both P_j and P_l choose to prioritize alignment (implicitly assuming successful coordination leads to $1((d_j, d_l) \in T_{jl}) = 1$ for these pairs).

As van den Steen (2017) notes, the underlying situation can be viewed as a coordination game with potentially multiple equilibria, especially if participants have differing views on which decision(s) to align *with*. Here, however, we abstract from these equilibrium selection issues and focus on the aggregate outcome based on individual incentives to align. This baseline represents the organization's performance without strategic guidance, where each participant independently resolves the trade-off between external fit and internal coherence based on their local information and judgment, assuming coordination occurs when alignment is chosen.

4.2 The Value of Strategic Control

We now analyze how much value a strategist creates when exercising formal authority over a specific decision. When a strategist has control over decision D_k, they directly determine whether that decision prioritizes stand-alone optimality or alignment with another decision. This intervention affects not only the target decision but potentially the entire pattern of alignments throughout the organization.

Proposition 1 (Value of Strategic Control)**.** *When strategist S has control over decision D_k, the strategic value created is:*

$$SV_{control}^{S}(D_k) = \begin{cases} \alpha_k \nu_k^S - \mathbb{I}_k^P \cdot \alpha_k \nu_k^P + \gamma \cdot (N_{align}^{S_k=1} - N_{align}^{base}) & \text{if } \mathbb{I}_k^S = 1 \\ \gamma \cdot (N_{align}^{S_k=0} - N_{align}^{base}) - \mathbb{I}_k^P \cdot \alpha_k \nu_k^P & \text{if } \mathbb{I}_k^S = 0 \end{cases} \tag{10}$$

Where:

- $\mathbb{I}_k^S = 1$ *if* $\alpha_k \nu_k^S \geq \gamma$ *(strategist prefers stand-alone optimization for* D_k*)*
- $\mathbb{I}_k^P = 1$ *if* $\alpha_k \nu_k^P \geq \gamma$ *(participant prefers stand-alone optimization for* D_k*)*
- $N_{align}^{S_k=1}$ *is the number of aligned pairs when strategist optimizes* D_k *for stand-alone correctness*
- $N_{align}^{S_k=0}$ *is the number of aligned pairs when strategist aligns* D_k *with another decision*

Proof. When the strategist controls decision D_k, the expected revenue is:

$$\mathbb{E}[R|\text{control } k] = \begin{cases} \alpha_k \nu_k^S + \sum_{j \neq k} S_j^{P|S_k=1} \cdot \alpha_j \nu_j^P + \gamma \cdot N_{\text{align}}^{S_k=1} & \text{if } \mathbb{I}_k^S = 1 \\ \sum_{j \neq k} S_j^{P|S_k=0} \cdot \alpha_j \nu_j^P + \gamma \cdot N_{\text{align}}^{S_k=0} & \text{if } \mathbb{I}_k^S = 0 \end{cases} \tag{11}$$

The strategic value is the difference between this value and the baseline:

$$\text{SV}_{\text{control}}^S(D_k) = \mathbb{E}[R|\text{control } k] - \mathbb{E}[R|\text{baseline}] \tag{12}$$

Note that other participants' stand-alone choices remain the same, as their decision rule depends only on their local parameters. The only change is in the alignment patterns. Substituting and simplifying yields the result. □

Given this, we can see that the value of strategic control increases with:

1. The strategist's confidence advantage $(\nu_k^S - \nu_k^P)$ when the strategist prefers stand-alone optimization
2. The extent to which control improves alignment patterns $(N_{\text{align}}^{S_k=x} - N_{\text{align}}^{\text{base}})$
3. The alignment importance γ when improved alignment is achieved

This proposition reveals that strategic control creates value through two distinct mechanisms:

1. **Direct decision quality effect:** When the strategist chooses stand-alone optimization, value is created if the strategist has higher confidence than the participant $(\nu_k^S > \nu_k^P)$. This represents the classic "better decisions" value of expertise.
2. **Coordination effect:** Control creates value by improving the overall pattern of alignments throughout the organization, as captured by the term $\gamma \cdot (N_{\text{align}}^{S_k=x} - N_{\text{align}}^{\text{base}})$. Importantly, this value can emerge even if the strategist has lower confidence about stand-alone correctness than the participant.

Example 1 (Strategic Control in a Retail Chain)**.** *Consider a retail chain with three key decisions: market selection (D_1), store format (D_2), and product assortment (D_3). Suppose that the alignment benefit γ likely exceeds the expected stand-alone value for* *store format, $\alpha_2\nu_2^P$. However, suppose that the participant erroneously prioritizes store format's stand-alone optimality over alignment with market selection. A strategist with control over store format decision who correctly recognizes that $\alpha_2\nu_2^S < \gamma$ creates value not through better judgment about store formats per se, but by improving the coordination between store formats and market selection. This coordination value can be substantial even if the strategist has lower expertise about store formats ($\nu_2^S < \nu_2^P$).*

For AI versus human strategists, this result has important implications. First, in domains where the primary value comes from improved decision quality (high α_k, significant confidence advantage), the strategist with higher ν_k^S creates more value through control. This means AI may excel in data-rich domains where it develops higher confidence, while humans maintain advantages in judgment-rich domains. By contrast, in domains where the primary value comes from improved coordination (high γ, significant alignment improvements), what matters is which strategist better recognizes when alignment should be prioritized over stand-alone optimality. Even a strategist with lower confidence about stand-alone correctness can create substantial value if they better understand the fundamental trade-offs between external fit and internal coherence.

4.3 The Value of Strategic Announcements

When strategists lack formal authority, they must rely on influence through strategic announcements. We now analyze the value created through such announcements and compare it to the value of control.[5]

Proposition 2 (Value of Strategic Announcements). *When strategist S makes a strategic announcement regarding decision D_k without formal control, the strategic value created is:*

$$SV^S_{announce}(D_k) = \rho_k^S \cdot SV^S_{control}(D_k) \cdot \frac{\nu_k^P}{\nu_k^S} \cdot \mathbb{I}_k^S + \rho_k^S \cdot SV^S_{control}(D_k) \cdot (1 - \mathbb{I}_k^S) \qquad (13)$$

[5] The examination of handing decision authority to AIs versus humans is not new and has been explored at the task level by Athey et al. (2020). What is new here is the focus on strategic leadership and influence in a broader system.

Or, alternatively:

$$SV^S_{announce}(D_k) = \begin{cases} \rho^S_k \cdot \left[\alpha_k \nu^P_k - \mathbb{I}^P_k \cdot \alpha_k \nu^P_k + \gamma \cdot (N^{S_k=1}_{align} - N^{base}_{align})\right] & \textit{if } \mathbb{I}^S_k = 1 \\ \rho^S_k \cdot \left[\gamma \cdot (N^{S_k=0}_{align} - N^{base}_{align}) - \mathbb{I}^P_k \cdot \alpha_k \nu^P_k\right] & \textit{if } \mathbb{I}^S_k = 0 \end{cases} \tag{14}$$

where ρ^S_k is the probability that participant P_k follows strategist S's recommendation.

Proof. When the strategist makes an announcement without control, the expected revenue is a weighted average of two scenarios:

$$\mathbb{E}[R|\text{announce } k] = \rho^S_k \cdot \mathbb{E}[R|P_k \text{ follows strategist}] + (1 - \rho^S_k) \cdot \mathbb{E}[R|\text{baseline}] \tag{15}$$

When the participant follows the strategist's recommendation, the resulting pattern matches what would happen under control, but with ν^P_k replacing ν^S_k for the stand-alone correctness probability (since it's the participant's confidence that determines the expected value of stand-alone correctness, not the strategist's). The strategic value calculation follows directly from this weighted average. □

From this proposition, it can be seen that the value of strategic announcements increases with the agreement probability ρ^S_k, the potential value under control SV^S_{control} and the participant's confidence ν^P_k relative to the strategist's confidence ν^S_k when the strategist recommends stand-alone optimization.

This proposition reveals several crucial insights about strategic influence. First, Agreement is critical. The value of strategic announcements is directly proportional to the agreement probability ρ^S_k. Without agreement, even potentially valuable strategic guidance has no effect. Second, the confidence ratio matters. When recommending stand-alone optimization, the value depends on the ratio $\frac{\nu^P_k}{\nu^S_k}$. This creates an important tension: a strategist with very high confidence ν^S_k might create less value through announcements than expected if the participant has much lower confidence ν^P_k.Finally, announcements create value through the same mechanisms as control (direct decision quality and coordination), but scaled by the agreement probability and adjusted for the confidence ratio.

Example 2 (Strategic Announcement in a Technology Firm)**.** *Consider a technology firm deciding on platform architecture (D_1), feature prioritization (D_2), and go-to-market strategy (D_3). An AI strategist develops high confidence ($\nu^{AI}_2 = 0.9$) that* *feature prioritization should optimize for stand-alone correctness rather than alignment. However, the development team has only moderate confidence in their feature judgments ($\nu_2^P = 0.6$).*

Even with a high agreement probability ($\rho_2^{AI} = 0.8$), the value created through announcement is:

$$SV_{announce}^{AI}(D_2) = 0.8 \cdot SV_{control}^{AI}(D_2) \cdot \frac{0.6}{0.9} = 0.53 \cdot SV_{control}^{AI}(D_2)$$

This is substantially less than what might be expected from the agreement probability alone, due to the confidence gap between the AI strategist and the development team.

For comparing AI versus human strategists, this proposition highlights the importance of not just analytical capabilities but also implementation effectiveness. In particular, human strategists typically achieve higher agreement probabilities ($\rho_k^H > \rho_k^{AI}$) through shared mental models and communication capabilities. This can give humans an advantage in strategic influence even when their analytical capabilities are inferior. However, AI strategists may develop much higher confidence than participants in data-rich domains, creating a potentially problematic confidence ratio that reduces the value of their announcements. In contrast, human strategists' confidence levels are often better calibrated with participants, leading to more favorable confidence ratios.

4.4 The Incremental Value of Control

Having characterized the value of both control and announcements, we now analyze when control provides significant incremental value beyond what could be achieved through influence alone. This helps determine when formal authority should be allocated to different strategist types.

Proposition 3 (Incremental Value of Control)**.** *The incremental value of control over announcement for strategist S regarding decision D_k is:*

$$\Delta_k^S = \begin{cases} (1-\rho_k^S) \cdot \left[\gamma \cdot (N_{align}^{S_k=1} - N_{align}^{base}) - \mathbb{I}_k^P \cdot \alpha_k \nu_k^P\right] + \alpha_k \nu_k^S - \rho_k^S \cdot \alpha_k \nu_k^P & \textit{if } \mathbb{I}_k^S = 1 \\ (1-\rho_k^S) \cdot \left[\gamma \cdot (N_{align}^{S_k=0} - N_{align}^{base}) - \mathbb{I}_k^P \cdot \alpha_k \nu_k^P\right] & \textit{if } \mathbb{I}_k^S = 0 \end{cases} \tag{16}$$

Proof. The incremental value of control is simply the difference between the value created under control and the value created through announcement:

$$\Delta_k^S = \mathrm{SV}_{\mathrm{control}}^S(D_k) - \mathrm{SV}_{\mathrm{announce}}^S(D_k) \tag{17}$$

Substituting the expressions from Propositions 1 and 2 and simplifying yields the result. □

This implies that the incremental value of control over announcement increases with the disagreement probability $(1-\rho_k^S)$, increases with the confidence advantage $(\nu_k^S - \rho_k^S \cdot \nu_k^P)$ when recommending stand-alone optimization, increases with the importance of the coordination effect $\gamma \cdot (N_{\mathrm{align}}^{S_k=x} - N_{\mathrm{align}}^{\mathrm{base}})$ and is zero when $\rho_k^S = 1$ and $\nu_k^S = \nu_k^P$ (perfect agreement and equal confidence)

This proposition reveals the fundamental tension in allocating control rights. Control is most valuable precisely when influence is least effective—when agreement is low and when there are significant differences in confidence levels. First, disagreement drives control value, The term $(1-\rho_k^S)$ scales most components of incremental control value, indicating that formal authority becomes more valuable as disagreement increases. This is something which we explore in depth in Section 5. Second, confidence advantage matters differently by decision type. For stand-alone optimization decisions, control's incremental value depends critically on the strategist's confidence advantage. For alignment decisions, this factor disappears, making the disagreement probability even more central. Finally, there is zero incremental value when there is perfect agreement. In the limiting case where $\rho_k^S = 1$ and $\nu_k^S = \nu_k^P$, control offers no additional value beyond what could be achieved through perfect influence.

Example 3 (Incremental Control Value in a Financial Institution)**.** *Consider a financial institution deciding on investment allocation (D_1), risk management protocols (D_2), and client segmentation (D_3). For risk management protocols suppose that an AI strategist has high confidence ($\nu_2^{AI} = 0.9$) and recommends stand-alone optimization while the risk manager has moderate confidence ($\nu_2^P = 0.7$) and also prefers stand-alone optimization, Let the agreement probability be moderate ($\rho_2^{AI} = 0.6$).*

Under these assumptions, the incremental value of control is:

$$\Delta_2^{AI} = 0.4 \cdot \left[\gamma \cdot (N_{align}^{S_2=1} - N_{align}^{base}) - \alpha_2 \cdot 0.7\right] + \alpha_2 \cdot 0.9 - 0.6 \cdot \alpha_2 \cdot 0.7 \tag{18}$$

Simplifying we have:

$$\Delta_2^{AI} = 0.4 \cdot \gamma \cdot (N_{align}^{S_2=1} - N_{align}^{base}) + 0.2 \cdot \alpha_2 \tag{19}$$

This shows that the incremental value has two components: one driven by the coordination effect (scaled by the disagreement probability) and another driven by the confidence advantage.

These results have important potential implications for the relative advantages of AI versus human strategists. First, AI strategists typically achieve lower agreement probabilities ($\rho_k^{AI} < \rho_k^H$), which increases the incremental value of giving them control. However, this higher incremental control value coincides with less organizational credibility; a counter-intuitive implication that we explore in detail in the next section. Second, in data-rich domains where AI develops significantly higher confidence than humans ($\nu_k^{AI} > \nu_k^H$), control allocation to AI becomes more valuable specifically for stand-alone optimization decisions. In judgment-rich domains, humans maintain this advantage. Finally, the incremental value of control for alignment decisions depends primarily on disagreement probability, not on confidence advantages. This creates a more nuanced comparison for decisions where alignment is preferred.

4.5 Strategic Advantage: AI versus Human Strategists

Our analysis thus far provides a foundation for directly comparing AI and human strategists across different decision contexts and intervention approaches. We now formalize the conditions under which each strategist type holds a comparative advantage.

Proposition 4 (Comparative Strategic Advantage)**.** *AI has a comparative advantage over human strategists under control for decision D_k when:*

$$\Delta SV_{control}(D_k) = SV_{control}^{AI}(D_k) - SV_{control}^{H}(D_k) > 0 \tag{20}$$

This occurs when (1) for stand-alone optimization decisions ($\mathbb{I}_k^{AI} = \mathbb{I}_k^H = 1$): $\alpha_k(\nu_k^{AI} - \nu_k^H) + \gamma \cdot (N_{align}^{AI,S_k=1} - N_{align}^{H,S_k=1}) > 0$; (2) for alignment decisions ($\mathbb{I}_k^{AI} = \mathbb{I}_k^H = 0$), $\gamma \cdot (N_{align}^{AI,S_k=0} - N_{align}^{H,S_k=0}) > 0$ and (3) for divergent trade-off judgments (e.g., $\mathbb{I}_k^{AI} = 1, \mathbb{I}_k^H = 0$), AI has advantage when its judgment about the fundamental stand-alone versus alignment trade-off leads to higher expected revenue.

Similarly, AI has a comparative advantage under announcement when:

$$\Delta SV_{announce}(D_k) = \rho_k^{AI} \cdot SV_{control}^{AI}(D_k) \cdot \frac{\nu_k^P}{\nu_k^{AI}} - \rho_k^H \cdot SV_{control}^H(D_k) \cdot \frac{\nu_k^P}{\nu_k^H} > 0 \tag{21}$$

for stand-alone optimization decisions, with a similar expression for alignment decisions.

This proposition highlights several key drivers of comparative advantage:

1. **Confidence advantage:** For stand-alone optimization decisions, higher confidence directly contributes to comparative advantage. AI typically achieves higher confidence in data-rich domains, while humans maintain advantages in judgment-rich domains.

2. **Coordination effectiveness:** Even with similar confidence levels, strategists can differ in their ability to improve overall alignment patterns across the organization. This coordination effectiveness depends on understanding interdependencies between decisions.

3. **Trade-off judgment:** Perhaps most fundamentally, strategists may differ in their judgment about the basic trade-off between stand-alone optimization and alignment for specific decisions. Getting this fundamental trade-off right is often more important than incremental improvements in confidence.

4. **Implementation effectiveness:** Under announcement rather than control, comparative advantage also depends critically on relative agreement probabilities and confidence ratios. A strategist with lower potential value under control may still have comparative advantage if they achieve much higher agreement.

Example 4 (Comparative Advantage in Product Development)**.** *Consider a product development context with decisions on target market segment (D_1), technical architecture (D_2), and pricing (D_3). For the technical architecture decision, both an AI and human strategist prefer stand-alone optimization ($\mathbb{I}_2^{AI} = \mathbb{I}_2^H = 1$). The AI has higher confidence ($\nu_2^{AI} = 0.85$ vs. $\nu_2^H = 0.7$) but achieves lower agreement probability ($\rho_2^{AI} = 0.6$ vs. $\rho_2^H = 0.9$).*

Under control, AI has an advantage of:

$$\Delta SV_{control}(D_2) = \alpha_2 \cdot (0.85 - 0.7) + \gamma \cdot (N_{align}^{AI,S_2=1} - N_{align}^{H,S_2=1}) = 0.15 \cdot \alpha_2 + \gamma \cdot \Delta N_{align} \tag{22}$$

Under announcement, the comparison becomes:

$$\Delta SV_{announce}(D_2) = 0.6 \cdot SV^{AI}_{control}(D_2) \cdot \frac{\nu^P_2}{\nu^{AI}_2} - 0.9 \cdot SV^{H}_{control}(D_2) \cdot \frac{\nu^P_2}{\nu^{H}_2} \qquad (23)$$

Even with AI's confidence advantage, the human strategist's superior agreement probability may give them the comparative advantage under announcement.

This analysis reveals that the comparative advantage of AI versus human strategists depends critically on the data-richness, fundamental uncertainty, and interdependence patterns of specific decisions, whether strategy is implemented through control or influence, how each strategist develops confidence in different decision domains and how effectively each strategist generates voluntary agreement from participants.

4.6 Competitive Interactions

The data-uncertainty divide between AI and human strategists takes on additional dimensions when we extend our analysis to competitive settings. van den Steen (2018b) explored the role of strategy, by his definition, in competitive settings. Thus, it is natural to explore how competitive interactions impact the results thus far. Specifically, how does the effectiveness of AI versus human strategists vary across different competitive environments? Under what conditions might an AI-led firm gain a competitive advantage over a human-led rival?

In competitive environments, the data richness-uncertainty trade-off affects not only internal decision alignment but also the strategic interaction with competitors. Based on our results to date, it is easy to see that in competitive settings, AI and human strategists exhibit different patterns of effectiveness depending on the nature of competition. AI strategists have a comparative advantage in competitive environments where data is rich, fundamental uncertainty is low and while there may be rapid competitive response cycles, there are also established competitive patterns. In established competitive patterns with rich historical data, AI can better predict competitor responses and formulate optimal counter-strategies. By contrast, in novel competitive situations with limited precedent, human intuition and judgment provide advantages in forming strategic expectations.

Moving beyond these general insights, van den Steen (2018b) observes that quantity-based competitive moves (like capacity expansion or product launches) appear more strategic than price-based moves. This observation relates directly to our data-uncertainty framework but reveals surprising implications for AI strategists that challenge conventional wisdom.

In quantity competition, AI strategists have advantages in commitment credibility due to their algorithmic, rule-based nature and excel at optimization of complex capacity and production decisions, while the non-zero-sum nature of quantity competition allows AI's commitment advantages to create strategic value. This latter point arises because, as van den Steen (2018b) shows, the value of strategy in competition is positive so long as there are non-zero-sum elements to competition between firms. In price competition, arguably, human strategists have advantages in developing product differentiation strategies that soften price competition. Human creativity enables escape from zero-sum price dynamics through novel positioning, and the creative aspect of differentiation involves fundamental uncertainty where humans maintain advantages.

These arguments suggest that, contrary to conventional expectations, AI strategists might excel in quantity competition precisely because their algorithmic nature serves as a credible commitment device. When AI announces a capacity decision, this announcement may carry greater credibility because the AI is perceived as less likely to deviate from established decision rules. In contrast, human strategists maintain advantages in the creative aspects of competition—particularly in developing novel product differentiation strategies that transform price competition into more favorable competitive terrains.

When firms led by different types of strategists compete against each other, the strategic dynamics may differ from those of competition between similar strategists. This has important implications for competitive intensity and market outcomes. Specifically, markets with mixed AI and human strategists might exhibit not just more intense competition but qualitatively different competitive dynamics. AI strategists may aggressively commit to quantity decisions (capacity expansion, production levels, inventory stocking), leveraging their algorithmic consistency to make credible commitments. Human strategists, meanwhile, might respond with creative differentiation strategies that transform the competitive landscape in ways that AI's historical data cannot fully capture.

This strategic asymmetry between AI and human approaches creates a dynamic where each strategist type focuses on dimensions where it has a comparative advantage, leading to competition that unfolds across multiple strategic dimensions simultaneously. The reduced potential for tacit coordination further intensifies this competition.

Interestingly, this suggests that industries with mixed AI and human strategists might actually see higher levels of investment in both capacity (driven by AI credible commitment) and product innovation (driven by human creativity) compared to industries dominated by either type alone. This could potentially create markets with both greater productive efficiency and more diverse product offerings, though potentially at the cost of reduced industry profitability.

5 The Control-Credibility Relationship

We noted earlier that AI strategists typically achieve lower agreement probabilities than their human counterparts. The direct effect of this suggests that the incremental value of giving AI's control over decisions is relatively high. However, this is strongest when there is lower credibility on the part of AI implying that those decisions are unlikely to be the core of a strategy. This reveals an important and perhaps counterintuitive relationship between strategic credibility and the need for formal control. Contrary to conventional wisdom, AI strategists may require less formal control in precisely those domains where they possess the greatest analytical advantages. This control-credibility relationship has significant implications for organizational design and the effective integration of AI into strategic processes.

5.1 An Inverse Relationship

Building on the formal results of Section 4, we can establish a more precise characterization of when formal control becomes necessary versus when influence through strategic announcements suffices.

Proposition 5 (Control-Credibility Relationship)**.** *The incremental value of control (Δ_k^S) decreases as the strategist's credibility increases along two dimensions:*

1. ***Agreement probability:*** $\frac{\partial \Delta_k^S}{\partial \rho_k^S} < 0$ *for all decision contexts*
2. ***Participant confidence calibration:*** *As* ν_k^P *approaches* ν_k^S*, the* $\alpha_k(\nu_k^S - \rho_k^S \nu_k^P)$ *term in* Δ_k^S *approaches zero for stand-alone optimization decisions*

Thus, in environments where a strategist achieves high agreement probability and where participant confidence aligns with strategist confidence, the value of formal control diminishes substantially.

Proof. From Proposition 3, recall that the incremental value of control is:

$$\Delta_k^S = \begin{cases} (1-\rho_k^S)\cdot\left[\gamma\cdot(N_{\text{align}}^{S_k=1} - N_{\text{align}}^{\text{base}}) - \mathbb{I}_k^P\cdot\alpha_k\nu_k^P\right] + \alpha_k\nu_k^S - \rho_k^S\cdot\alpha_k\nu_k^P & \text{if } \mathbb{I}_k^S = 1 \\ (1-\rho_k^S)\cdot\left[\gamma\cdot(N_{\text{align}}^{S_k=0} - N_{\text{align}}^{\text{base}}) - \mathbb{I}_k^P\cdot\alpha_k\nu_k^P\right] & \text{if } \mathbb{I}_k^S = 0 \end{cases} \tag{24}$$

Taking the partial derivative with respect to ρ_k^S:

$$\frac{\partial\Delta_k^S}{\partial\rho_k^S} = \begin{cases} -\left[\gamma\cdot(N_{\text{align}}^{S_k=1} - N_{\text{align}}^{\text{base}}) - \mathbb{I}_k^P\cdot\alpha_k\nu_k^P\right] - \alpha_k\nu_k^P & \text{if } \mathbb{I}_k^S = 1 \\ -\left[\gamma\cdot(N_{\text{align}}^{S_k=0} - N_{\text{align}}^{\text{base}}) - \mathbb{I}_k^P\cdot\alpha_k\nu_k^P\right] & \text{if } \mathbb{I}_k^S = 0 \end{cases} \tag{25}$$

In contexts where strategic intervention creates value, both $[\gamma\cdot(N_{\text{align}}^{S_k=x} - N_{\text{align}}^{\text{base}}) - \mathbb{I}_k^P\cdot\alpha_k\nu_k^P]$ and $\alpha_k\nu_k^P$ are positive, making $\frac{\partial\Delta_k^S}{\partial\rho_k^S} < 0$.

For the second part, as ν_k^P approaches ν_k^S, the term $\alpha_k\nu_k^S - \rho_k^S\cdot\alpha_k\nu_k^P$ approaches $\alpha_k\nu_k^S(1-\rho_k^S)$, which approaches zero as ρ_k^S approaches 1. □

This proposition establishes that as a strategist becomes more credible—meaning participants increasingly agree with their strategic recommendations and develop confidence levels aligned with the strategist's—the incremental benefit of granting that strategist formal control diminishes. In the limiting case where $\rho_k^S = 1$ and $\nu_k^P = \nu_k^S$, formal control offers no additional value beyond what could be achieved through strategic influence alone.

5.2 AI Credibility Across Decision Domains

How does this control-credibility relationship specifically apply to AI versus human strategists? The answer lies in understanding how credibility varies systematically across different decision contexts.

Note first that AI strategists' credibility relative to human strategists varies systematically across the spectrum from data-rich to judgment-rich domains:

1. In data-rich domains (abundant historical data, well-defined metrics):

- $\rho_k^{AI} > \rho_k^H$ (higher agreement probability)
- $\frac{\nu_k^P}{\nu_k^{AI}} \approx \frac{\nu_k^P}{\nu_k^H}$ (similar confidence calibration)

2. In judgment-rich domains (limited historical data, tacit knowledge requirements):

   - $\rho_k^{AI} < \rho_k^H$ (lower agreement probability)
   - $\frac{\nu_k^P}{\nu_k^{AI}} < \frac{\nu_k^P}{\nu_k^H}$ (worse confidence calibration)

These domain-specific credibility patterns create distinct implications for control allocation. Specifically, the incremental value of control for AI versus human strategists shows an inverse relationship with data richness. That is, as decisions become more data-rich and less dependent on judgment, the comparative control premium for AI relative to human strategists decreases and may become negative.

This challenges the conventional wisdom that AI requires more formal control than human strategists to be effective. In fact, our analysis reveals that:

1. **Data-Rich Domains:** AI strategists typically achieve higher agreement probabilities ($\rho_k^{AI} > \rho_k^H$) due to their demonstrable analytical capabilities. This higher credibility means AI can create substantial value through influence without requiring formal control. Counterintuitively, human strategists may need more formal control than AI in these domains to overcome their relative credibility deficit.

2. **Judgment-Rich Domains:** AI strategists face lower agreement probabilities ($\rho_k^{AI} < \rho_k^H$) due to the subjective, experiential nature of these decisions. In these contexts, AI typically requires more formal control than human strategists to create comparable value, aligning with conventional expectations.

Thus, empirically wemay observe AI strategists having less control than their human counterparts.

Example 5 (Supply Chain Optimization versus New Market Entry)**.** *Consider two strategic decisions facing a manufacturing firm:*

Supply Chain Optimization (Data-Rich): *For this decision, the AI achieves high agreement (*$\rho_1^{AI} = 0.85$*) due to its transparent analysis of historical data and clear* *demonstration of efficiency improvements. In contrast, the human strategist achieves lower agreement ($\rho_1^H = 0.7$) due to perceived biases toward familiar suppliers. The incremental value of control is:*

$$\Delta_1^{AI} = 0.15 \cdot [\textit{coordination value}] + \alpha_1 \cdot (0.9 - 0.85 \cdot 0.8) = 0.15 \cdot [\textit{coordination value}] + 0.22 \cdot \alpha_1 \tag{26}$$

$$\Delta_1^{H} = 0.3 \cdot [\textit{coordination value}] + \alpha_1 \cdot (0.75 - 0.7 \cdot 0.8) = 0.3 \cdot [\textit{coordination value}] + 0.19 \cdot \alpha_1 \tag{27}$$

New Market Entry (Judgment-Rich): *For this decision, the AI achieves lower agreement ($\rho_2^{AI} = 0.4$) due to limited historical precedent and challenges in modeling complex competitive dynamics. The human strategist achieves higher agreement ($\rho_2^H = 0.8$) through narrative communication and industry experience. The incremental value of control is:*

$$\Delta_2^{AI} = 0.6 \cdot [\textit{coordination value}] + \alpha_2 \cdot (0.7 - 0.4 \cdot 0.6) = 0.6 \cdot [\textit{coordination value}] + 0.46 \cdot \alpha_2 \tag{28}$$

$$\Delta_2^{H} = 0.2 \cdot [\textit{coordination value}] + \alpha_2 \cdot (0.8 - 0.8 \cdot 0.6) = 0.2 \cdot [\textit{coordination value}] + 0.16 \cdot \alpha_2 \tag{29}$$

This illustrates that for supply chain optimization (data-rich), the incremental value of control might actually be higher for the human strategist, while for new market entry (judgment-rich), control creates substantially more incremental value for the AI strategist.

5.3 The Data-Control Trade-off

The inverse relationship between data richness and the need for control can be formalized more precisely by examining how key parameters influencing the incremental value of control (Δ_k^{AI}, defined in Proposition 3) change with data availability.

To see this, assume that for an AI strategist:

- Agreement probability, ρ_k^{AI}(data richness), is non-decreasing with data richness (or strictly increasing over some range).

- The AI's confidence, ν_k^{AI}(data richness), is non-decreasing with data richness (or strictly increasing over some range).
- The participant's confidence ν_k^P and preference $\mathbb{I}_k^P$ are relatively stable with respect to the data richness available to the AI strategist.

Under these assumptions, the incremental value of control for the AI strategist, Δ_k^{AI}(data richness), typically exhibits a non-monotonic relationship with data richness, potentially showing an inverted U-shape. It tends to be low for very low data richness (where AI confidence ν_k^{AI} is low, limiting the potential value creation from control) and low for very high data richness (where agreement ρ_k^{AI} approaches 1, reducing the gap between control and influence), peaking at intermediate levels.

Recall from Proposition 3 (Equation 17) that Δ_k^{AI} depends positively on the disagreement probability $(1-\rho_k^{AI})$ and, when $\mathbb{I}_k^{AI}=1$, on the confidence advantage term involving $(\nu_k^{AI}-\rho_k^{AI}\nu_k^P)$. It also depends on the coordination improvement effect, scaled by $(1-\rho_k^{AI})$. Consider how these components change with data richness under the stated assumptions:

1. As data richness increases, the AI's confidence ν_k^{AI} tends to increase. This generally increases the potential value generated by control (e.g., through the $\alpha_k\nu_k^{AI}$ term if $\mathbb{I}_k^{AI}=1$), potentially making control more valuable, ceteris paribus.
2. As data richness increases, the agreement probability ρ_k^{AI} tends to increase. This *decreases* the disagreement factor $(1-\rho_k^{AI})$, which scales most components of Δ_k^{AI}, thereby reducing the incremental value of control relative to influence.

The overall shape of Δ_k^{AI}(data richness) depends on the interplay between these opposing forces. For very low data richness, ν_k^{AI} is likely low, meaning the AI adds little value even with control, making Δ_k^{AI} small. For very high data richness, ρ_k^{AI} approaches 1, meaning $(1-\rho_k^{AI})$ approaches 0. Even if ν_k^{AI} is high, influence becomes nearly as effective as control, making Δ_k^{AI} small again. Therefore, the incremental value of control Δ_k^{AI} typically peaks at intermediate levels of data richness where the AI has gained sufficient capability (ν_k^{AI} is significant) but hasn't yet achieved near-perfect agreement (ρ_k^{AI} is still significantly below 1). The exact shape depends on the specific functional forms of $\nu_k^{AI}(\cdot)$ and $\rho_k^{AI}(\cdot)$.

This data-control trade-off has profound implications for AI integration into strategic processes:

1. **Low Data Environments:** In domains with minimal structured data, AI offers limited strategic value (low ν_k^{AI}) regardless of control allocation. These domains are better suited to human strategists who can leverage tacit knowledge and pattern recognition.

2. **Moderate Data Environments:** In domains with moderate data availability, AI can identify valuable strategic patterns (ν_k^{AI} is significant) but may struggle to communicate them convincingly or achieve high agreement (ρ_k^{AI} is moderate). These "transition domains" typically represent the peak of the control need curve (Δ_k^{AI} is highest), where formal authority enables AI to overcome credibility limitations while still creating substantial value.

3. **Rich Data Environments:** In domains with abundant, high-quality data, AI can both identify optimal strategies (ν_k^{AI} is high) and convincingly demonstrate their value, leading to high agreement (ρ_k^{AI} approaches 1). These domains require minimal formal control for AI to be effective (Δ_k^{AI} is low); influence through transparent analysis often suffices.

5.4 Organizational Design Implications

Our analysis reveals a fundamental factor for organizations integrating AI into strategic processes: in data-rich domains where AI demonstrates superior analytical capabilities, formal control becomes less necessary as agreement naturally emerges. This relationship suggests that a more sophisticated approach than simply replacing human strategists with AI systems is required.

In particular, it is arguable that the optimal allocation of formal control to AI versus human strategists follows a domain-contingent pattern:

1. **Judgment-Rich Domains:** Human strategists with advisory AI support

2. **Transition Domains:** AI strategists with formal control but human oversight

3. **Data-Rich Domains:** AI strategists operating primarily through influence, with humans maintaining formal control for accountability

Consider a global retailer facing strategic decisions across its enterprise. For data-dense inventory optimization, AI can operate effectively through influence alone, providing transparent analyses that managers readily accept due to demonstrable performance advantages. For market entry decisions involving ambiguous competitive dynamics, human strategists maintain natural advantages through narrative reasoning and contextual understanding. In ”transition domains” like production technology selection—where data exists but patterns remain contested—a hybrid approach emerges as optimal, with AI granted formal decision authority but operating under human oversight.

To implement this domain-contingent approach, organizations would have to consider developing three key mechanisms:

1. **Differentiated Authority Systems:** Explicitly distinguish between formal control rights and influence channels across decision domains, implementing decision-specific authority allocations rather than blanket AI authority or purely advisory roles.

2. **Progressive Control Models:** As AI demonstrates credibility in specific domains, its formal control needs typically diminish. Organizations should implement models where AI initially receives formal control to overcome credibility limitations, then transitions to influence-based roles as agreement probabilities increase.

3. **Credibility Enhancement Mechanisms:** Invest in systems that enhance AI credibility without requiring formal control, including transparent reasoning processes that explain AI strategic recommendations, track record documentation that builds credibility through demonstrated success, and hybrid communication approaches where human interpreters contextualize AI analysis.

The organizations that will perform well in this new strategic landscape are those that design explicit mechanisms to enhance AI credibility where it’s weakest while leveraging human judgment where it’s strongest. This nuanced integration recognizes that strategic success depends not simply on analytical excellence but equally on the credibility necessary to transform insights into coordinated action.

5.5 Synthesizing the Control-Credibility Relationship

The key insight from our analysis is that the relationship between strategic credibility and control need is not merely correlational but causal: higher credibility directly reduces the need for formal control by enabling effective influence-based implementation.

For AI strategists, this creates an important dynamic:

1. **Initial Control Premium:** When first deployed in a decision domain, AI may require a ”control premium”—formal authority that exceeds what might be justified by its analytical capabilities alone. This premium compensates for initially low agreement probabilities.

2. **Self-Diminishing Control Need:** As AI demonstrates effectiveness in a domain, the very success that justifies its strategic role also reduces its need for formal control. Successful AI strategists essentially work themselves out of formal authority positions through progressive credibility building.

3. **Domain-Specific Evolution:** Different decision domains will evolve along different trajectories. Data-rich domains may quickly transition to influence-based AI roles, while judgment-rich domains may require sustained formal control or remain better suited to human strategists.

This nuanced understanding replaces the simplistic narrative that AI requires more control than humans to be effective strategists. Instead, it reveals that control needs depend on specific domain characteristics, credibility patterns, and organizational learning processes. The optimal approach is neither complete AI autonomy nor rigid human oversight, but rather a dynamic, domain-specific allocation of formal authority that evolves as credibility relationships mature.

By recognizing and managing this control-credibility relationship explicitly, organizations can more effectively integrate AI into their strategic processes, capturing analytical advantages while avoiding unnecessary centralization of authority. Most importantly, this approach acknowledges that strategic effectiveness depends not only on identifying the optimal stand-alone versus alignment trade-offs, but also on generating the organizational credibility needed to implement these insights effectively.

6 Conclusion

This paper has examined the conditions under which artificial intelligence can effectively perform the role of a strategist, extending (van den Steen, 2017, 2018a) formal theory of strategy. Our analysis reveals that the fundamental question is not simply whether AI can formulate strategy, but rather how organizations must transform when AI assumes strategic functions. The answer to each question yields markedly different insights about the future of strategic leadership.

The implications of our analysis extend far beyond the simple substitution of human strategists with AI systems. In a manufacturing firm, for instance, an AI strategist might prioritize supply chain optimization decisions as core strategic elements due to its superior pattern recognition in data-rich contexts, while a human strategist might focus on brand positioning decisions where judgment and narrative construction are paramount. This difference in strategic emphasis would necessitate not just different decisions being designated as "strategic," but fundamentally different organizational structures to support implementation.

Perhaps most intriguingly, AI strategists offer the possibility of unprecedented transparency in strategic thinking. While human CEOs face inherent limitations in bandwidth—unable to simultaneously meet with the head of marketing, operations, and R&D to explain strategic rationale—an AI strategist could simultaneously engage with multiple stakeholders, providing consistent explanations of its strategic reasoning without the cognitive constraints human strategists face. Consider how this might transform a retail chain implementing a new market strategy: rather than cascading communication through hierarchical layers with inevitable distortion, all regional managers could simultaneously engage with the AI strategist to understand the precise strategic logic, improving both buy-in and implementation fidelity. It is this type of system-wide change that Agrawal et al. (2022) argue will be at the core of making AI transformative.

The "control paradox" identified in our analysis—where AI strategists require less formal control in domains where they possess the greatest analytical advantages—further challenges conventional thinking about authority in organizations. A pharmaceutical company employing an AI strategist might maintain human control over early-stage R&D decisions where judgment about scientific novelty is paramount, while granting the AI considerable autonomy in clinical trial design and analysis where its pattern recognition excels. This domain-contingent approach to control allocation represents a significant departure from traditional organizational design principles.

While our analysis has entered the realm of what might currently be considered science fiction, it offers a deeper appreciation of organizational design imperatives that have always existed but rarely been articulated: organizations need to adapt to the capabilities of their strategists as much as strategists must match their organizations. This resonates with the finding in van den Steen (2018a) that strategy formulation by a CEO yields better execution than formulation by an outsider or consultant, precisely because the CEO's control over implementation provides credibility. Our analysis suggests a similar dynamic applies to AI strategists; like consultants, they may offer superior data analysis but lack the inherent credibility stemming from implementation control that human strategists (akin to CEOs) possess, thus requiring different organizational adaptations. The literature has extensively explored how strategists should align with organizational context, but has given less attention to how organizations might be deliberately designed to match the specific strengths and limitations of different types of strategists. Furthermore, the emphasis on data-rich domains often aligns AI's strengths with more stable, mature industries, suggesting a potential niche where AI strategists might initially prove most effective.

As AI capabilities continue to advance, these theoretical insights will take on increasing practical relevance. Organizations that develop sophisticated mechanisms for AI-human strategic collaboration—differentiating authority systems across decision domains, implementing progressive control models as AI credibility develops, and investing in credibility enhancement mechanisms—will likely outperform those that either resist AI involvement in strategy or attempt wholesale replacement of human strategic judgment.

The future of strategic leadership lies not in choosing between human intuition and AI analysis, but in designing organizational structures that effectively integrate both, recognizing that strategic effectiveness ultimately depends on the alignment between a strategist's capabilities and the organization's design. This bidirectional relationship between strategist and organization represents fertile ground for future research that extends beyond the traditional boundaries of strategic management theory.

References

Agrawal, A., Gans, J., and Goldfarb, A. (2018). *Prediction machines: The simple economics of artificial intelligence*. Harvard Business Press.

Agrawal, A., Gans, J., and Goldfarb, A. (2022). *Power and prediction: The disruptive economics of artificial intelligence*. Harvard Business Press.

Agrawal, A., Gans, J. S., and Goldfarb, A. (2024). Artificial intelligence adoption and system-wide change. *Journal of Economics & Management Strategy*, 33(2):327–337.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*.

Athey, S. C., Bryan, K. A., and Gans, J. S. (2020). The allocation of decision authority to human and artificial intelligence. In *AEA Papers and Proceedings*, volume 110, pages 80–84. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Brynjolfsson, E. and McAfee, A. (2017). The business of artificial intelligence. *Harvard Business Review*, 7:3–11.

Brynjolfsson, E. and Rock, D. (2022). The ai productivity paradox. *Journal of Economic Perspectives*, 36(3):3–24.

Chalmers, D., MacKenzie, N. G., and Carter, S. (2021). Artificial intelligence and entrepreneurship: Implications for venture creation in the fourth industrial revolution. *Entrepreneurship Theory and Practice*, 45(5):1028–1053.

Ehrig, T. and Schmidt, J. (2022). Theory-based learning and experimentation: How strategists can systematically generate knowledge at the edge between the known and the unknown. *Strategic Management Journal*, 43(7):1287–1318.

Felin, T., Gambardella, A., and Zenger, T. (2024). Theory-based decisions: Foundations and introduction. *Strategy Science*, 9(4):297–310.

Felin, T. and Zenger, T. R. (2017). The theory-based view: Economic actors as theorists. *Strategy Science*, 2(4):258–271.

Gans, J. S. (2024). Internal disagreement and disruptive technologies. *Strategy Science*, 9(3):267–276.

Griffith, T. L. (1999). Technology features as triggers for sensemaking. *Academy of Management review*, 24(3):472–488.

Hambrick, D. C. (2007). Upper echelons theory: An update. *Academy of Management Review*, 32(2):334–343.

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-ai symbiosis in organizational decision making. *Business Horizons*, 61(4):577–586.

Kahneman, D. (2011). *Thinking, Fast and Slow.* Farrar, Straus and Giroux, New York.

Keating, G. (2012). *Netflixed: The epic battle for America's eyeballs.* Penguin.

Klein, G. (2007). Naturalistic decision making. *Human Factors*, 49(5):970–986.

Knight, F. H. (1921). *Risk, Uncertainty and Profit.* Houghton Mifflin, Boston, MA.

Lovallo, D. and Kahneman, D. (2003). Delusions of success: How optimism undermines executives' decisions. *Harvard Business Review*, 81(7):56–63.

Morris, S. (1995). The common prior assumption in economic theory. *Economics and Philosophy*, 11(2):227–253.

Raisch, S. and Krakowski, S. (2021). Artificial intelligence and management: The automation-augmentation paradox. *Academy of Management Review*, 46(1):192–210.

Shrestha, Y. R., Ben-Menahem, S. M., and von Krogh, G. (2019). Organizational decision-making structures in the age of artificial intelligence. *California Management Review*, 61(4):66–83.

Tichy, N. M. and Sherman, S. (1993). *Control Your Destiny or Someone Else Will: How Jack Welch Is Making General Electric the World's Most Competitive Corporation.* Doubleday, New York.

van den Steen, E. (2010a). Interpersonal authority in a theory of the firm. *American Economic Review*, 100(1):466–490.

van den Steen, E. (2010b). On the origin of shared beliefs (and corporate culture). *The RAND Journal of Economics*, 41(4):617–648.

van den Steen, E. (2017). A formal theory of strategy. *Management Science*, 63(8):2616–2636.

van den Steen, E. (2018a). Strategy and the strategist: How it matters who develops the strategy. *Management Science*, 64(10):4533–4551.

van den Steen, E. (2018b). The strategy in competitive interactions. *Strategy Science*, 3(4):574–591.

von Krogh, G. (2018). Artificial intelligence in organizations: New opportunities for phenomenon-based theorizing. *Academy of Management Discoveries*, 4(4):404–409.

Wuebker, R., Zenger, T., and Felin, T. (2023). The theory-based view: Entrepreneurial microfoundations, resources, and choices. *Strategic Management Journal*, 44(12):2922–2949.

Table 1: Model Parameters and Variables

Symbol	**Description**
Exogenous Parameters	
K	Number of interdependent decisions in the project.
$\mathcal{D}_k$	Set of possible choices for decision D_k.
T_k	Unknown true state representing the objectively correct choice for D_k in isolation.
S	Strategist
P	Participant
α_k	Economic importance of decision D_k being correct on its own (stand-alone correctness).
T_{kl}	Interaction state defining the alignment requirement (bijection) between D_k and D_l.
γ_{kl}	Economic importance of achieving alignment between decisions D_k and D_l.
γ	Simplified uniform interaction importance ($\gamma_{kl} = \gamma$) used for analysis.
ν_k^P	Participant P_k's subjective confidence that their belief θ_k^P matches T_k.
ν_k^S	Strategist S's confidence that their belief θ_k^S matches T_k.
$\nu_k^{S,H}$	Confidence of a human strategist.
$\nu_k^{S,AI}$	Confidence of an AI strategist.
c_S	Cost for strategist S to investigate a decision state $T_{\bar{k}}$.
ρ_k	Agreement parameter: probability $\theta_k^P = \theta_k^S$.
ρ_k^H	Agreement parameter for a human strategist.
ρ_k^{AI}	Agreement parameter for an AI strategist.
Endogenous Variables	
D_k	Decision k within the project.
d_k	Actual choice made for decision D_k.
R	Total project revenue, depends on all d_k relative to T_k and T_{kl}.
R_k	Revenue contribution associated with decision D_k.
θ_k^P	Participant P_k's belief (best estimate) about the state T_k.
θ_k^S	Strategist S's belief (best estimate) about the state T_k after investigation.
M	Strategy announcement (message) made by the strategist S.
λ_k	Control indicator: 1 if strategist S controls decision D_k, 0 otherwise.
S_k	Indicator function: 1 if decision k prioritizes stand-alone correctness.
A_{kl}	Indicator function: 1 if decision k is deliberately aligned with decision l.