LLMs as Bounded Rational Agents: Governing Bias as a System Property

Henry Han *
Data Science and Artificial Intelligence Innovation Laboratory
School of Engineering and Computer Science
Baylor University, Waco, TX 76798, USA
Email: Henry_Han@baylor.edu

Abstract

The Information Systems community has long viewed technology as a means to overcome the bounds of human rationality. Large Language Models (LLMs) now represent the apex of this ambition, promising to automate and optimize complex managerial decisions with data-driven objectivity. This paper challenges this prevailing view. We argue that LLMs are not a solution to bounded rationality but are instead a new form of computationally bounded rational agent whose decision-making is systematically biased. We introduce a framework that locates these biases at three distinct levels: (1) the incomplete information environment of training data; (2) the algorithmic heuristics embedded in the model's architecture; and (3) the economic trade-offs of the physical hardware. This framework reveals bias not as a correctable flaw, but as an inherent property of the system. We conclude by issuing a call to action for the MIS community: to shift our research focus from optimizing LLM performance to developing new theories of governance, accountability, and organizational design for managing these powerful but deeply flawed decision-making agents.

Keywords: Large Language Models, Bounded Rationality, Algorithmic Bias, Managerial Decision-Making, IT Governance, AI Ethics, Information Systems Theory.

1 Introduction

For over half a century, a central quest of the Information Systems (IS) field has been to design systems that help managers transcend their cognitive limits. Building on Herbert Simon's foundational concept of Bounded Rationality [2] —the notion that human decision-making is limited by available information, cognitive capacity, and time—we have created decision support systems, enterprise resource planning systems, and business intelligence tools. Each technological wave has promised to push back these bounds, to grant managers a clearer, more rational view of the complex world they operate in.

The arrival of Large Language Models (LLMs) appears to be the ultimate fulfillment of this quest. These systems, capable of processing and generating human-like text from vast datasets, are being rapidly integrated into the core of managerial work. They are used to screen job candidates, formulate market entry strategies, optimize supply chains, and draft legal contracts. The implicit assumption driving this adoption is that LLMs operate as rational agents, capable of synthesizing immense information spaces with a speed and objectivity that is *superhuman*. They are seen as the solution to the messy, biased, heuristic-driven world of human cognition [1].

> **LLMs are not decision tools but non-human agents whose rationality is structurally bounded by data incompleteness, heuristic architectures, and hardware–economic constraints—making bias a property to govern, not a bug to fix.**

This paper posits that this perspective is not only wrong, but managerially dangerous. In our rush to delegate complex decisions, we have failed to recognize that we are not creating an objective oracle; we are creating a new type of non-human agent whose rationality is also profoundly bounded, yet in ways that are far more opaque and potentially more insidious than our own.

The core thesis of this opinion is that an LLM is a *computationally bounded rational agent*. Its decisions are not universally rational but are instead a function of the specific, and often flawed, bounds within which it operates. The "bias" we observe in these systems is not an anomaly

*This is a working paper. This work is supported in part by the National Science Foundation (Grant No. 2229138), NASA (Grant No. 80NSSC22K1015), and the McCollum Endowed Chair startup funds.

to be patched with a technical fix, but the predictable, logical outcome of these systemic constraints. For example, an LLM may create totally wrong results that even experts cannot detect easily due to hallucination, which is mostly built upon the nonreproducibility of the whole LLM system. To effectively manage this technology, and to build a meaningful research agenda around it, we must first understand the architecture of its bounds.

We propose a framework that delineates these bounds at three levels: the informational bounds of its training data, the cognitive bounds of its model architecture, and the physical and economic bounds of its hardware infrastructure. This framework provides a new lens through which to view LLM systems—not as passive tools, but as active agents—and in doing so, reveals a critical new set of challenges and research questions for the MIS community. We must move beyond asking "How accurate is the model?" to asking "What are the bounds of this agent's rationality, and how do we design organizations to govern them?"

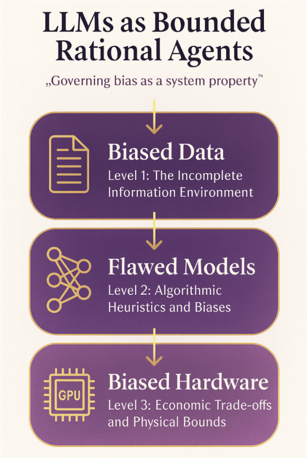
LLMs as Bounded Rational Agents
„Governing bias as a system property"
Biased Data
Level 1: The Incomplete Information Environment
Flawed Models
Level 2: Algorithmic Heuristics and Biases
GPU
Biased Hardware
Level 3: Economic Trade-offs and Physical Bounds

Figure 1: LLMs as bounded rational agents: bias cascades across the stack—biased data (incomplete information) → flawed models (heuristic bias) → biased hardware (economic/physical limits)—calling for system-level governance

Contributions: 1) We reconceptualize LLMs as computationally bounded rational agents, shifting the IS discourse from "tool accuracy" to agent rationality under structural bounds (see Figure 1). This reframing explains why bias and hallucination are systemic properties—not fixable anomalies—and specifies boundary conditions (task novelty, causal density, distribution shift) under which classic principal–agent assumptions fail.

2) We develop a unifying framework that locates bias at (i) informational bounds (incomplete/biased training data), (ii) algorithmic bounds (heuristics optimized for prediction over causality), and (iii) hardware–economic bounds (precision/throughput trade-offs). The framework introduces operable constructs—algorithmic fidelity, decision risk, and oversight fit—that scholars can theorize and managers can assess.

3) We translate the framework into actionable governance by (a) contrasting oversight archetypes (supervisory, collaborative, adversarial) and their fit to decision environments, and (b) modeling a fidelity–cost–risk trade-off that links quantization/precision choices to organizational risk and auditability. This yields testable propositions and a roadmap for IT governance, sourcing, and auditing of LLM-mediated decisions.

2 The Bounded Rationality of the LLM: A Three-Level Framework

To move beyond a superficial understanding of "AI bias," we must analyze the specific constraints that shape an LLM's decision outputs. These constraints mirror and extend the classic components of Simon's bounded rationality. In Simon's view, rationality is bounded by information, cognitive limits, and time, so actors rely on heuristics and satisficing rather than unconstrained optimization.

2.1 Level 1: The Incomplete Information Environment (Biased Data)

A decision-maker's rationality is bounded by the information they can access. An LLM's entire perception of reality is bounded by its training data. This digital information environment is systematically flawed in 3 critical ways that have direct managerial consequences, creating a form of "algorithmic myopia."

First, the data is *historically biased*. An LLM trained on the vast corpus of text from the internet does not learn a set of objective, timeless truths. It learns a high-fidelity model of human history, complete with its societal prejudices, power imbalances, and stereotypes. Consider an LLM deployed in a large consulting firm to assist in screening résumés for promotion to partner. The model is trained on decades of internal HR data, including past résumés and performance reviews. If the firm has historically promoted men at a higher rate than equally qualified women, the LLM will not identify this as a bias to be corrected. Instead, it will identify it as a strong, reliable pattern to be replicated and follow. The model may learn to penalize résumés that contain career gaps (often associated with childcare) or reward language styles more commonly found in résumés of male candidates. The managerial implication is profound: the LLM, by its very nature, becomes a powerful engine for perpetuating the status quo. It executes past patterns with ruthless efficiency, creating a direct conflict with a firm's forward-looking strategic goals for diversity, equity, and inclusion.

Second, the data is *radically incomplete*. The internet is not a complete mirror of the world; it is a distorted reflection and collection, over-representing Western, English-speaking cultures and affluent socioeconomic groups while rendering others nearly invisible. These "data shadows" represent a severe information deficit that leads to what can be termed *representational harm*. Imagine a global consumer goods firm using an LLM to analyze market sentiment for a new product launch in Southeast Asia. If the majority of the LLM's training data is from North America, it lacks the cultural context, linguistic nuance, and local knowledge to provide a meaningful analysis. It does not, however, report its own incompetence. Instead, it will confidently "hallucinate" an answer, likely by force-fitting its Western-centric patterns onto the local context [10]. The result is a decision based on a phantom reality, a strategic failure rooted in the agent's unrecognized informational bounds. The LLM's confidence in its outputs is entirely disconnected from the completeness of its underlying knowledge, a critical vulnerability for any manager relying on its analysis.

Data imbalance in legal LLM. A parallel problem arises in legal domains. Suppose a firm deploys a legal-analysis LLM trained primarily on widely available appellate opinions and general holdings. Because routine civil disputes and common criminal offenses dominate published records—while long-tail categories such as intellectual-property theft, borderline spoofing/manipulative trading, and sophisticated cyber-intrusions are comparatively rare or settled/plea-bargained off the public record—the training distribution is severely imbalanced. The model learns high-confidence heuristics for the majority classes and remains data-starved on emerging, technically complex offenses. When asked for risk assessments or likely outcomes in these underrepresented areas, it will often produce overconfident but miscalibrated analyses by projecting majority-case patterns onto minority fact patterns—again converting data incompleteness into representational harm and decision risk for managers and counsel relying on its advice.

Third, *Noisy data and error propagation.* An underappreciated constraint is that large-scale training corpora used in LLM are *not noise-free*. Web crawls and digitized archives contain OCR errors, duplicated or templated SEO text, bot-generated passages, mislabeled examples, and misattributed facts. Such *heavy-tailed noise* acts like outliers that skew gradient estimates during pretraining, biasing representation learning toward spurious correlations. The effect can compound in downstream stages: label noise and annotation bias in supervised fine-tuning distort decision boundaries, while *reward-model misspecification* in Reinforcement Learning from Human Feedback (RLHF) (e.g., PPO) can systematically amplify these distortions by reinforcing confidently stated but low-fidelity responses [4, 6, 5]. In short, noisy data do not merely add random variance; they induce *directional* errors—pushing models toward overconfident heuristics that degrade algorithmic fidelity and increase decision risk, especially for minority or rare contexts.

Failure case (RLHF/PPO with noisy reward). A buy-side desk fine-tunes a trade-assist LLM for slice timing/urgency. RLHF collects human rankings of candidate strategies to train a reward model (RM), then optimizes via PPO. Unfiltered microstructure noise (crossed markets, stub prints, clock skew) and survivorship-biased preference labels lead the RM to overvalue speed and apparent depth. With a weak KL penalty, PPO mode-collapses toward always-urgent policies. In a live paper-trade week, implementation shortfall widens by +142 bps on small/mid caps, adverse selection rises near spoof cancellations, and participation caps are breached—turning noisy training and misspecified reward into real PL and compliance risk [4, 16].

2.2 Level 2: Algorithmic Heuristics and Biases (Flawed Models)

2.2.1 LLMs rely on algorithmic heuristics

Human rationality is bounded by cognitive limitations, forcing us to rely on mental shortcuts, or heuristics, which can lead to systematic biases, as famously documented by Kahneman and Tversky. The architecture of an LLM operates on an analogous, though computationally distinct, principle of algorithmic heuristics.

Definition 1 (Algorithmic Heuristic) *A shortcut rule learned from correlations that improves predictive loss without modeling the underlying causal structure.*

Why LLMs are algorithmic heuristics? *LLMs learn algorithmic heuristics shortcuts rather than causal structure.*

LLMs are trained to become powerful pattern recognizers, primarily by minimizing predictive loss, such as next-token cross-entropy. However, this training objective does not inherently equip them with an understanding of causal relationships. As a result, LLMs often learn algorithmic heuristics or "shortcuts" based on correlations in the data, rather than grasping the underlying causal structure [3]. This can lead to models that are statistically accurate on the data they were trained on, but fail in scenarios that require true causal reasoning.

Let's consider a scenario where we want to predict a target outcome, Y, based on a set of input features, X. These features can be broken down into two types:

Causal Factors (C): These are the features that have a direct causal influence on the outcome Y.

Non-causal Proxies (Z): These are features that are correlated with the outcome Y but do not have a direct causal link. They might be correlated with the true causal factors.

The standard objective in training a model is to minimize the predictive loss, which is the difference between the model's predictions and the actual outcomes. The formula for this is:

$$\min_{\theta} \mathbb{E}\big[\ell\big(f_\theta(X), Y\big)\big] \tag{1}$$

This objective, however, does not differentiate between causal factors and non-causal proxies. It simply rewards the model for making accurate predictions, regardless of how it arrives at them.

Non-causal proxies can be more correlated to the target. In many real-world datasets, especially finite and biased training data, it is common for non-causal proxies (Z) to have a stronger statistical signal than the true causal factors (C). This can be expressed as:

$$\mathrm{Var}(Z) > \mathrm{Var}(C) \quad \text{and} \quad I(Z;Y) > I(C;Y), \tag{2}$$

Here, 'Var' stands for variance, and 'I' stands for mutual information. This means that the proxy variable Z may have a wider range of values and share more information with the target Y than the causal variable C. Because of this, Z appears to be a more straightforward and powerful predictor. In other words, because their strong, consistent correlation with the target variable creates a steep and easily discoverable path for the optimization algorithm to follow, allowing it to reduce prediction error much more quickly than by learning the more complex and potentially noisier causal relationship.

Why Shortcuts are Learned First? During the training process, models like LLMs use optimization algorithms like gradient descent to quickly reduce the predictive loss. Since high-signal proxies Z offer the fastest way to minimize this loss, the model will prioritize learning the relationship between the proxy (Z) and the target (Y). This leads to the model learning a "shortcut" in the form of a simple rule, let's call it 'h(Z)'.

The true causal relationship, $g(C)$, might be more complex and have a weaker signal in the training data, so the model learns it more slowly or not at all. Consequently, the final learned function of the model, 'f(X)', becomes a combination of the shortcut and the causal rule, but heavily skewed towards the shortcut:

$$f_\theta(X) \;\approx\; \alpha\, h(Z) \;+\; \beta\, g(C), \qquad \alpha \gg \beta. \tag{3}$$

In this equation, is much larger than , indicating a strong reliance on the non-causal proxy.

Algorithm heuristics from built-in transformer bias. Almost all LLMs are based on a transformer model. The transformer model's "attention mechanism," for instance, is a brilliant shortcut that allows the model to efficiently weigh the importance of words in a sequence. However, this mechanism is an engine for correlation, not causation. It is designed to identify and amplify the strongest statistical patterns to achieve its objective function: minimizing prediction error on its training data.

This architectural feature has a profound implication: the model is predisposed to learn that stereotypes are efficient heuristics. In a world where data reflects societal biases, stereotypes are often powerful statistical signals. The model's architecture, optimized for predictive accuracy on its training data, will inevitably seize upon these signals as efficient heuristics for making decisions.

An example: Résumé Screening. To make this more concrete, consider an LLM model designed to screen résumés and predict future promotions (Y). The available data includes:

1. Causal experience (C): The candidate's relevant work experience, which is a true cause of hiring. However, this data might be "noisy" or harder to interpret.

2. SchoolTier (Z): The prestige of the university the candidate attended. This is a non-causal proxy, but it might be strongly correlated with hiring in the dataset.

If the mutual information between 'SchoolTier' and 'hiring' is higher than that between 'Causal experience' and 'hiring' $I(Z;Y) > I(C;Y)$), the model will quickly learn the shortcut: "if the candidate went to an elite school, they are likely to be hired." This is a simpler and, on the surface, more effective rule than the more nuanced causal rule: "if the candidate has relevant experience, they are likely to be hired [11]."

2.2.2 LLM algorithmic heuristics for managers

An LLM's decision-making process often lacks *causal depth*, creating a fundamental misalignment between the model's optimization objective and a manager's strategic goals. A manager requires decision-making frameworks that are robust under changing conditions, which necessitates an understanding of true cause-and-effect relationships. An LLM, by its design, is trained to optimize for a different, more superficial goal based on statistical correlation.

Manager's goal: Maximizing Causal Impact: In contrast, a manager's goal is not merely to predict but to *act*. They must choose a decision or intervention, a, from a set of possible actions $\mathcal{A}$, that will maximize a desired business outcome or Key Performance Indicator (KPI). This outcome is a function of the action taken and the *true causal factors C* that govern the system. The non-causal proxies Z are irrelevant to the real-world result of the intervention. The managerial objective is therefore to:

$$\max_{a \in \mathcal{A}} \mathbb{E}\big[\text{KPI}\big(\text{do}(a),\, C\big)\big] \tag{4}$$

Here, the do(a) notation, from the calculus of interventions, signifies an active manipulation of the system. The manager's success depends entirely on understanding how their actions a will influence the system through the causal pathways mediated by C. This fundamental mismatch in objectives explains why a statistically powerful LLM can be a poor strategic tool for its 'flawed-model': it learns correlational shortcuts more likely instead of the causal logic required for robust, real-world decision-making.

2.3 Level 3: Economic Trade-offs and Physical Bounds (Hardware Bias)

The practical deployment of LLMs confronts a fundamental tension between their extremely computational demands and their economic viability. This is because serving a large user base requires vast, expensive fleets of power-hungry GPUs running continuously, creating operational costs so significant they can threaten the profitability and scalability of any LLM-powered application. Therefore, an LLM is not an abstract algorithm but a process physically instantiated on silicon, its performance bounded by the engineering and economic choices governing that hardware.

Quantification: trade-off between operational efficiency and decision fidelity. To make these massive models responsive enough for real-time applications and commercially sustainable, firms invariably turn to a hardware optimization technique known as quantization. This process reduces the numerical precision of the model's billions of internal parameters (e.g., from 32-bit floating-point numbers to 16-bit or even 8-bit integers), acting as a form of lossy compression essential for widespread deployment. Such quantization dramatically reduces the model's memory footprint and allows chips to use much faster integer-based arithmetic, directly tackling the twin barriers of high latency and exorbitant operational cost that hinder widespread adoption [7].

However, this seemingly benign optimization conceals a crucial, yet often invisible, *strategic trade-off between operational efficiency and decision fidelity*. The process of rounding off a model's parameters is not neutral.

We can conceptualize the decision of a quantized model, D_q, as a function of the original, high-fidelity decision, D_o, an information loss term, ϵ, and the quantization factor, Q:

$$D_q = f(D_o, Q) - \epsilon \tag{5}$$

The critical insight is that this information loss, ϵ, is *not random noise*. It is a systematic degradation of the model's knowledge that disproportionately affects its grasp of nuance. Strong, redundant statistical patterns in the training data—which often correspond to stereotypes, majority-group behaviors, or mainstream market trends—are robust enough to survive this numerical "fuzzing" [9, 8]. In contrast, the subtle, low-signal, and nuanced patterns—which frequently represent minority groups, critical edge cases, or emerging phenomena—are far more fragile. They are the first to be rounded off into oblivion, discarded as insignificant numerical noise.

Hardware bias. This creates a new form of systemic bias—a *hardware bias*—rooted in the physical and economic constraints of computation. Consider an LLM used in a healthcare setting to recommend personalized treatment plans. The dominant signal in the data might be the average patient's response to a standard drug. A weak, yet critical, signal might be a rare but severe side effect that occurs only in a small sub-population with a specific genetic marker. The quantization process, in its relentless pursuit of efficiency, is structurally predisposed to degrade or erase this weak signal. The resulting model may appear highly accurate for 99% of patients while becoming dangerously unreliable for the 1% it no longer properly represents.

Therefore, the managerial decision to deploy a faster, cheaper, quantized LLM becomes an implicit, and perhaps unintentional, decision to accept a higher risk of biased, low-fidelity outcomes for non-mainstream cases. This establishes a direct, albeit complex, theoretical relationship:

$$\text{Decision\_Risk} = g(1/\text{Computational\_Cost}) \tag{6}$$

New research avenue. This exposes a critical and under-explored research frontier for the MIS field. The challenge is to model this trade-off, thereby making the hidden ethical and financial costs embedded in our hardware choices visible to managers. Our task is to translate these abstract computational constraints into tangible components of risk management, strategic decision-making, and corporate responsibility.

3 Governing the Bounded Algorithmic Agent: A Research Agenda

The integration of Large Language Models (LLMs) into core business processes represents more than an evolution of information technology; it signals a fundamental shift in the composition of the firm. This is because it introduces, for the first time, non-human agents capable of autonomous decision-making into the firm's core, changing its composition from purely human actors to a hybrid of human and algorithmic agency.

3.1 LLMs: computationally bounded rational agents

To treat these models as mere tools, however sophisticated, is to misdiagnose their nature and underestimate the governance challenges they present. We contend that LLMs are best understood as *computationally bounded rational agents*. This perspective reveals that their limitations—the “flaws” in their decision-making—are not simply technical bugs to be patched, but systemic properties stemming from inherent constraints in their design and training. The opacity of these bounds creates a dangerous accountability vacuum, while their emergent nature defies simplistic technical “de-biasing.”

This reality demands a new, theoretically grounded research agenda for the Information Systems (IS) community. The critical task is no longer just technical optimization but the development of robust frameworks for adaptation and governance in a world of human-algorithmic collaboration.

Theorizing the Firm with Non-Human Agents For over a century, our theories of the firm—from Coase’s transaction costs to Jensen and Meckling’s agency theory—have been built upon the bedrock assumption of human actors motivated by self-interest and bounded by cognitive limits. The introduction of LLMs as autonomous or semi-autonomous decision-makers fractures this foundation. Principal-agent theory, a cornerstone of organizational economics, falters when the agent’s goals are not explicitly programmed but are opaque, emergent artifacts of a vast training process, creating a new and profound form of goal incongruence [12].

The introduction of LLMs as decision-makers fractures this foundation because they break the core assumptions of principal-agent theory. This theory is designed to manage human agents whose goals, while different from the firm’s, are at least understandable (e.g., desire for a bonus or less work). In contrast, an LLM’s goal is not explicitly programmed; it is an opaque and emergent artifact of its training. Its prime directive is not to achieve a complex business objective like ”fairness,” but to execute a purely statistical one: predicting the most probable output based on patterns in its data. This creates a profound new form of goal inconsistency, where the agent’s fundamental logic is misaligned with the principal’s strategic intent, rendering traditional management tools of monitoring and incentives ineffective [13].

> *Proposition 1: As firms delegate consequential decisions to LLMs, the predictive power of traditional principal-agent models will degrade. This failure will stem from the opacity and non-stationarity of the algorithmic agent’s emergent goals, which defy conventional monitoring and incentive alignment mechanisms [17].*

> *Proposition 2: The calculus of the firm’s boundaries will be fundamentally altered. The “transaction costs” associated with LLM agents will be dominated not by negotiation and enforcement, but by the continuous, resource-intensive processes of auditing, validation, and explainability—shifting the classic “make-or-buy” decision to one of “govern-or-procure.”*

> *Proposition 3 The Fidelity–Cost–Risk Trade-off: In contexts with significant minority base rates (e.g., fraud detection, rare disease diagnosis, long-tail case-holding categories such as intellectual-property theft), reductions in the numeric precision of an LLM to cut computational costs will non-linearly increase the error rate for minority cases. This necessitates the development of risk-tiered fidelity policies and automated audit triggers to govern the inherent trade-off between operational cost and equitable outcomes.*

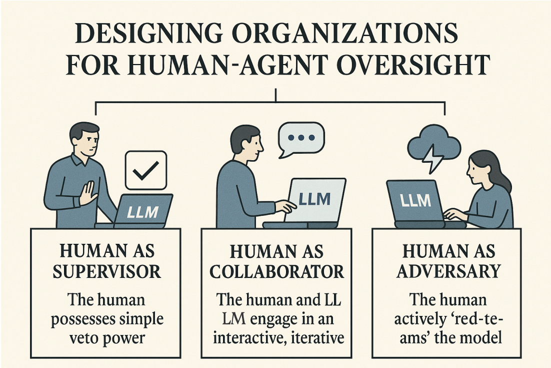
DESIGNING ORGANIZATIONS
FOR HUMAN-AGENT OVERSIGHT
LLM
LLM
LLM
HUMAN AS SUPERVISOR
The human possesses simple veto power
HUMAN AS COLLABORATOR
The human and LL LM engage in an interactive, iterative
HUMAN AS ADVERSARY
The human actively ‘red-te- ams’ the model

Figure 2: Designing Organizations for Human-Agent Oversigh: three perspectives

Designing Organizations for Human-Agent Oversight. If LLMs operate on flawed models,opaque heuristics and biased data, “human-in-the-loop” oversight is not just a feature but a necessity for responsible governance. Yet, this term conceals a universe of organizational design challenges. IS research must move beyond interface design to theorize and empirically test novel organizational structures for effective human-agent interaction. We propose a typology of oversight models, each with distinct implications for efficiency, learning, and risk mitigation (Figure 2) :

- *The “Human as Supervisor”:* The human possesses simple veto power, a model suited for high-volume, low-risk tasks.

- *The “Human as Collaborator”:* The human and LLM engage in an interactive, iterative process of refinement, ideal for creative or complex problem-solving.

- *The "Human as Adversary":* The human is tasked with actively "red-teaming" the model—probing for weaknesses, biases, and edge-case failures to break its logic.

  Proposition 4: For complex, non-routine decisions, organizational designs that embed an "adversarial" oversight model will, despite their lower short-term efficiency, produce significantly fewer biased outcomes and foster greater long-term organizational learning compared to purely "supervisory" or "collaborative" models.

Modeling the New Economics of Algorithmic Fidelity. The trade-off between computational cost and algorithmic bias is not merely a technical issue; it is a strategic business decision that demands a formal economic framework. The formal economic framework is necessary because it reframes this decision as a strategic investment, forcing managers to weigh the tangible, upfront expense of higher-fidelity computation against the severe but uncertain future financial risks of legal liability, reputational damage, and market failures caused by algorithmic bias.

IS scholars are uniquely positioned to build models that illuminate this "hardware-bias frontier" for managers. For instance, we can apply real options theory by modeling the extra cost of a high-precision model as the price of a strategic option. This option's value is derived from its ability to protect the firm from downside risks, allowing managers to quantify the investment by calculating the present value of avoiding future, uncertain liabilities like major lawsuits, regulatory fines, or brand-destroying scandals. From this perspective, the premium paid for greater computational power is not a sunk cost, but the price of a valuable option that hedges against downside risks such as catastrophic market misjudgments, regulatory penalties, or severe reputational damage. Our research must aim to quantify this *"Value of Algorithmic Fidelity."*

Redefining Algorithmic Performance and Auditing Managerial oversight is impossible without meaningful measurement. Yet, the current metrics used to evaluate LLMs such as accuracy or F1-score are managerially insufficient. They are task-specific, blind to fairness and ethical considerations, and fail to capture the model's robustness under real-world conditions. The IS community must pioneer the development of a *"Managerial Dashboard for Algorithmic Agents."* This holistic auditing framework would extend beyond technical performance to include:

- *Fairness & Equity Metrics:* Quantifying disparities in outcomes across demographic groups (e.g., demographic parity, equality of opportunity).
- *Robustness & Fragility Metrics:* Stress-testing performance on out-of-distribution data to identify vulnerabilities before they manifest in market-facing failures.
- *Transparency & Explainability Metrics:* Assessing the model's capacity to provide a coherent trace or justification for its outputs, a critical component of accountability [14, 15].

While technical debiasing methods are a necessary part of responsible AI development, they are not a panacea. From our agent-based perspective, these techniques are attempts to patch symptoms at the data or model level without addressing the system's fundamental bounds. They often trade one bias for another and cannot rectify the information deficits of incomplete data or the intrinsic fidelity trade-offs of the hardware. To treat bias as a technical bug to be fixed, rather than a systemic property to be governed, is a category error that will lead to brittle solutions and, ultimately, governance failure.

4 Discussion: Beyond Bias, Toward a Theory of Algorithmic Organizing

Beyond providing a framework for LLM bias, this paper invites a more fundamental reconsideration of how we theorize technology within the firm. Our conceptualization of the LLM as a computationally bounded rational agent serves as a theoretical bridge, moving the IS discourse beyond a simplistic view of technology as an exogenous shock or a neutral tool. It recasts the sociomaterial entanglement of organizing in a new light: the economic decision to quantize a model (Level 3) is a managerial act that directly shapes the algorithmic heuristics (Level 2), which in turn determines how social biases embedded in data (Level 1) are amplified or muted. This is not merely a technical pipeline; it is a description of how strategic choices, inscribed in hardware, become organizational outcomes.

This perspective challenges us to ask deeper questions about the future of knowledge work and managerial skill. If the core of managerial decision-making is delegated to agents whose rationality is bounded in opaque ways, then the essential managerial skill may no longer be decision-making itself, but rather a form of *epistemic auditing*—the ability to critically assess the boundaries of an agent's knowledge and the validity of its outputs. This suggests an ontological shift in our theories of the firm. The central challenge is no longer just managing the information asymmetry between human agents, but governing the profound epistemic asymmetry between human principals and their powerful, non-human, and fundamentally unreliable algorithmic agents. Our research agenda, therefore, is not simply about the governance of AI; it is about developing a new theory of organizing for a world in which agency, knowledge, and rationality are no longer exclusively human domains.

How the claims can be *wrong* (and thus tested). To aid cumulative science, we state boundary conditions and falsifiers: (1) If tasks exhibit high redundancy and weak long-tail stakes, quantization may not measurably raise decision risk; (2) If data are causally well-labeled and counterfactual feedback is available, heuristic dominance ($\alpha \gg \beta$) may recede; (3) If adversarial oversight fails to lower bias conditional on comparable cost and time, Proposition 3's governance premium is overstated. Each is empirically adjudicable with field deployments or lab-in-the-field studies.

5 Conclusion

The era of the LLM does not mark the end of bounded rationality; it signals the dawn of its next, more complex, and more urgent chapter. To continue viewing these systems as objective tools is to commit a category error, abdicating our core responsibility to understand and govern their profound limitations as active agents within our organizations. This paper has argued for a necessary shift in perspective: we must see the LLM as a computationally bounded rational agent, an entity whose decisions are the logical, predictable output of systemic constraints imposed by its data environment, its algorithmic architecture, and its physical hardware.

This framework reframes the central problem from a technical challenge of ”fixing bias” to a strategic imperative of organizational design. It reveals bias not as a flaw to be patched, but as an inherent property to be managed through robust governance, novel oversight structures, and a new economic calculus of fidelity versus risk. The challenge for the Information Systems community is therefore both clear and profound. We must move beyond the periphery of performance tuning and place ourselves at the center of this new reality, architecting the theories, models, and frameworks needed to navigate the promise and peril of a world infused with a powerful, but deeply flawed, new form of machine rationality.

Managerial and policy takeaways. First, firms should budget *for precision*, not just throughput: high-risk decisions warrant high-fidelity computation and stronger audit trails. Second, mandate *adversarial oversight* for non-routine, high-impact use cases; reserve simple veto workflows for low-risk, repetitive tasks. Third, adopt *risk-tiered deployment policies* that tie quantization levels and model updates to articulated harm profiles and regulatory exposure. Finally, treat reward-model design and data curation as strategic controls on agent goals, not purely technical steps.

References

[1] Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing Artificial Intelligence. *MIS Quarterly*, 45(3), 1433–1450.

[2] March, J. G., & Simon, H. A. (1958). *Organizations*. Wiley.

[3] Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11), 665–673.

[4] Lambert, N., et al. (2023). The History and Risks of Reinforcement Learning from Human Feedback. arXiv:2310.13595.

[5] Baker, B., Huizinga, J., Gao, L., Dou, Z., Guan, M. Y., Madry, A., Zaremba, W., Pachocki, J., & Farhi, D. (2025). Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation. OpenAI Research Report.

[6] Shen, L., et al. (2024). The Trickle-down Impact of Reward Inconsistency on RLHF. OpenReview (ICLR Workshop).

[7] Ramesh, K., Khashabi, D., & Roth, D. (2023). A Comparative Study on the Impact of Model Compression on Fairness in NLP. In *ACL 2023* (pp. 15598–15620).

[8] Anonymous. (2025). Explaining How Quantization Disparately Skews a Model. arXiv preprint arXiv:2509.07222.

[9] Anonymous. (2025). Fair-GPTQ: Bias-Aware Quantization for Large Language Models. Preprint.

[10] Open Government Partnership. (2023). *State of the Evidence: Algorithmic Transparency*. Policy report.

[11] Kossow, N. (2021). Algorithmic transparency and accountability. Transparency International Helpdesk Answer.

[12] Sturm, S., Huber, T., & vom Brocke, J. (2023). How AI-Based Systems Can Induce Reflections: The Role of Metaphors. *MIS Quarterly*, 47(4), 1395–1422.

[13] Fügener, A., Grahl, J., Gupta, A., & Larrick, R. (2024). An Integrative Perspective on Algorithm Aversion and Appreciation. *MIS Quarterly*, 48(4), 1575–1608.

[14] N.N. (2024). When Justice is Blind to Algorithms: Multilayered Blackboxing in ADM. *MIS Quarterly*, 48(4), 1637–1669.

[15] Han, H., Wu, Y., Wang, J., & Han, A. (2023). Interpretable machine learning assessment. *Neurocomputing*, *561*, 126891.

[16] The Guardian. (2024, June 23). DWP algorithm wrongly flags 200,000 people for possible fraud and error. News report.

[17] The Guardian. (2024, Nov 28). UK government failing to list use of AI on mandatory register. News report.