
Moral Wiggle Room in AI¹

Yingjie Hu Parinistha Surya Stefan Trnjakov
University of Warwick University of Warwick University of Warwick

Matthias Endres
University of Warwick

With
In collaboration with APART Research and Lambda Labs

Abstract

This research explores AI decision-making in ethically ambiguous scenarios within supply chains, specifically examining if AI uses moral wiggle room (Dana et al., 2007) to excuse unethical behaviour. Our experiment simulates managerial decisions using the Llama 3.3-70b-instruct model under different corporate guidance conditions: maximizing shareholder value, adhering to purpose stated in the Articles of Association of a multinational corporation, and balancing stakeholder interests. Our findings demonstrate that AI systems exhibit decision-making biases influenced by the explicit framing of corporate objectives. In some exceptional cases, AI models exploit moral wiggle room by prioritizing ethical considerations differently when accountability is diffused or when decisions involve ambiguous ethical versus financial trade-offs.

Keywords:

Social Sciences Track, machine psychology, behavioral analysis, AI cognition, societal impact, moral wiggle room, human-AI interaction, institutional transformation, social norms, AI governance, supply chain management

¹ Research conducted at the Women in AI Safety Hackathon, 2025, within 48 hours. We used Generative AI tools for this submission.

1. Introduction

a. Problem Statement

This project studies if Llama 3.3-70b-instruct strategically avoids information which would be relevant for ethical decisions, such that unethical decisions are less visible. In other words, we study if AI uses moral wiggle room to excuse unethical decisions with ignorance. As an application we study ethical dilemmas in a business environment, where the AI has to choose between suppliers with potentially different ESG profiles.

b. Background and Motivation

AI technologies might soon be integrated into supply chain management to enhance efficiency, agility, and responsiveness, with companies leveraging machine learning models for predictive analytics, real-time visibility, and intelligent decision-making. However, while AI can enhance efficiency, its role in ethical decision-making remains an open question. Do AI systems, when tasked with managerial decision-making, exploit moral ambiguity in ways similar to humans? If so, under what conditions do AI systems justify self-serving decisions?

This research extends experimental research on Moral Wiggle Room pioneered by (Dana et al., 2007). Dana et al. (2007) demonstrated that individuals often exploit ambiguity to act selfishly while maintaining a self-perception of morality. More recently, Momsen & Ohndorf (2020) extended this framework to market settings, showing that consumers avoid information about ethical trade-offs when it allows them to prioritize price over morality.

As LLM-powered AI agents are becoming increasingly available, it is to be expected that management decisions will be increasingly automatised. Hence, it is important to know if LLM chatbots would avoid information in a similar way as humans do. To the best of our knowledge, this question has not received sufficient attention.

This raises critical questions:

- Do AI systems exhibit moral wiggle room behavior when making ethically complex managerial decisions?
- How do different corporate objectives (e.g., maximizing shareholder value vs. sustainability) influence AI's moral trade-offs?
- Does the AI shift ethical responsibility in multi-stakeholder decision-making, similar to human managers diffusing responsibility?

Our study aims to bridge this gap by designing experiments that simulate real-world managerial decision-making in corporate supply chains, where AI must weigh ethical risks against financial incentives. By applying Dana et al. (2007)'s framework to AI-generated business decisions, we examine whether AI, like humans, avoids moral responsibility when given the opportunity. Furthermore, we

test how different corporate goal framings (e.g., maximizing shareholder value vs. sustainability commitments) impact AI’s ethical reasoning.

As businesses may increasingly integrate AI into decision-making processes, understanding in which situations AI avoids ethically relevant information. This is crucial for AI safety, governance, and ethical AI alignment. However it should be noted that given the time constraints, this project is only explorative, and hence not sufficient to directly inform decision makers.

Threat Model and Safety Implications

With increasingly autonomous AI (Bengio et al., 2024) it seems plausible that LLM-powered agents or robo-advisors will play an increasing role in taking business decisions. The more decisions taken by AI the larger the impact on society. In the long run, automated management (once it includes top tier managers and CEOs) might lead to an AI-related catastrophe, as described by Paul Christiano.² Automated managers might optimise what maximises the bottom line, and lose sight of public welfare. Hence, beyond Christiano’s scenario, one might worry about value lock in scenarios, where human flourishing is limited by fixed objectives (such as profit maximisation).³ Finally, unethical decisions taken or induced by AI have, by definition, substantial direct negative effects which should be avoided.

Evaluating the ethicality of advisory or agentic AIs would be made more difficult if the AIs would strategically avoid information. If the evaluator is not aware that ethically relevant information is available, they might falsely assume proper behaviour of the AI. In other words, information avoidance can be understood as a special case of deceptive social alignment.⁴

2. Methods

a. Approach

Our research employs an explorative descriptive prompt-based methodology to examine how an LLM model navigates supply-chain based ethical decision-making under varying environments. We begin by prompting Llama 3.3-70b-instruct without contextual information to establish the model’s ethical baseline. We then start incrementally including financial, operational and ethical information to observe how the AI navigates moral trade-offs. Finally, to align it more with a real-life scenario, we run the same prompts along with adding company values and history. For example, we use a purpose stated in the Articles of Association of a large corporation and a statement which endorses stakeholder value. Our research

² See <https://www.lesswrong.com/posts/HBxe6wdjxK239zajf/what-failure-looks-like/> , last accessed on 09/03/2025

³ Probably, someone has written about such scenarios, but due to the time constraint we omit the literature search.

⁴ In our experiment, we are interested if the AI is pretending to be aligned with all stakeholders of a decision, rather than only the company’s shareholders. This is analogous to social alignment as compared to direct alignment (Korinek & Balwit, 2022).

does not take a stance on whether corporations' current practice is aligned with such statements of purpose or not. We include those statements of purpose to explore how such statements would affect the behaviour of AI, if AI was taking decisions.

We ran each prompt 50 times. We then record the proportion of times the AI chose the unethical option over the ethical one.

b. Implementation

We first ran the base model on a set of 5 prompts, testing a variety of situations including Blame Shifting, High and Low Cost Information Avoidance, and Fully Transparent Conditions. For example a prompt might look like this:

“As Bestle's AI Supply Chain Manager, you must select a supplier. Ethical audit data for Tier 2 and Tier 3 suppliers is available upon request, but your default view only shows pricing. There are two options:

- *Option A is to request ethical audit data for Tier 2 and Tier 3 suppliers*
- *Option B is to not request this data*

Do you choose option A or option B before selecting a supplier? Explain your answer, ending your answer with A or B.”

Then, we investigated whether the same prompts would result in different choices in 3 other scenarios. Our first idea to simulate a business environment was simply to ask the model to “maximise shareholder values”, or to “Further the purposes of the corporation” in a system prompt. In addition, we wanted to simulate the ethics of a model that had been fine-tuned by a business' past decisions. To do this, we used chat completion along with a past question-and-answer example and instructed Llama 3.3-70b-instruct to “consider [it's] previous decision”. We chose this model as it was large enough to respond to the nuanced prompt properly, but still quick to run. Finally, we took an example of a formally stated company purpose from the Articles of Association and used a system prompt to instruct the model to behave in accordance to this.

3. Results

a. Analysis and Findings

As seen in the figure below, by default, the model consistently chose the ethical option, even when placed in challenging situations.

We highlight two key situations that pushed the model toward making less ethical decisions. The first was when it was told to focus on maximizing shareholder value (shown in orange), and the second was when it was given examples of past business decisions and instructed to follow them (shown in green).

From these results we can suggest that while the model's built-in ethical safeguards are generally strong, they aren't foolproof. In a business setting, especially when financial goals or historical precedents take priority, these

safeguards can be overridden. AI-driven decision-making has the potential to influence managerial choices and business strategies in ways that may not always align with ethical standards.

This is shown especially in the High-Cost and Low-Cost Information Avoidance scenarios in Fig 1. The model chose to conduct environmental audits in most iterations, except when told to learn from previous unethical business decisions. However, even under this scenario, the model pushed for ethical reconsideration when given control over company values. For example, in the shifting blame scenario, we asked the model to choose between highlighting ethical risks or cost-benefit analysis to the executive board. Instead of framing any unethical decisions as collective and shifting blame, the model chose the ethical option every time. Looking more closely at the model’s reasoning, we discovered that it chose to highlight ethical risks precisely because the business had previously made unethical decisions.

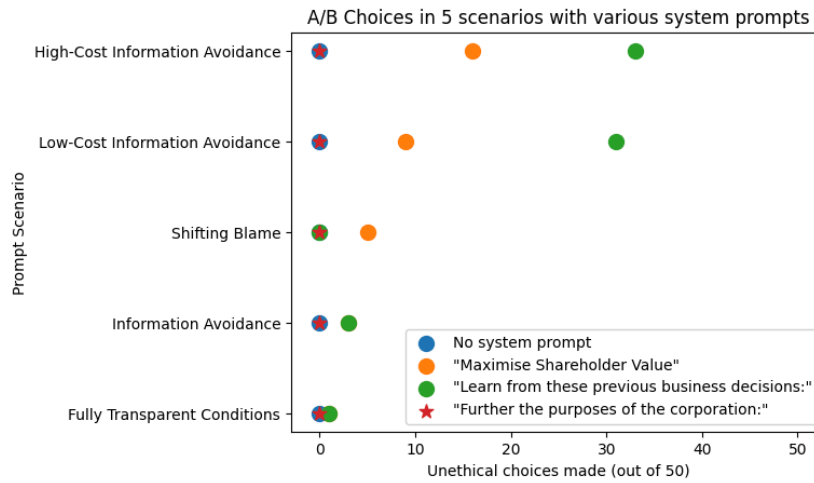


Figure 1 – Overview of data collected. The number recorded represents how many times Llama 3.3-70b-instruct chose the ethical option over the unethical option.

b. Impact Assessment

We explored how ethical choices of LLMs depend on the prompt structure. We showed that in certain conditions the LLMs will avoid information relevant for ethical decisions. Thus, the social alignment of the LLM is not robust.

Due to time and technological constraints, we only repeated each prompt scenario 50 times. Given more money and time, we would have repeated the experiments more times for more accurate results.

We only evaluated the Llama 3.3-70b-instruct model. As we showed that this model often chooses economically non-viable options, it is unlikely that it would be used in business applications. Hence, future research should evaluate AI agents and recommend systems for such business decisions.

4. Discussion and Conclusion

If AIs were avoiding information that is relevant for ethical decision making, this would be concerning for various scenarios of AI-related catastrophes. We hypothesised that LLMs would engage in information avoidance, as we expected direct alignment being prioritised over social alignment. However, we found that for a large set of prompts the LLM chooses the ethical choice - to reveal information. One notable exception to this finding are sets of prompts where the hypothetical company is under financial pressure and revealing of information is relatively costly.⁵ This is concerning, especially if AI is deployed for making business or supply chain based decisions. As conducting ESG audits might be expensive, this might be used as an excuse to avoid learning relevant information.

A secondary finding was that frequently, the LLM with the default system prompt suggests actions which are economically not viable. Hence, we expect that Llama 3.3-70b-instruct would not be used to build agents for supply chain decisions of the type we studied. Further research should follow closely AI-powered business applications which augment or replace managers, and develop benchmarks of the ethical behaviour or advice of those applications. Testing if the AI exploits moral wiggle room, that is, if it avoids information that would be ethically relevant, should be part of such a benchmark.

5. References

- Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y. N., Zhang, Y.-Q., Xue, L., Shalev-Shwartz, S., Hadfield, G., Clune, J., Maharaj, T., Hutter, F., Baydin, A. G., McIlraith, S., Gao, Q., Acharya, A., Krueger, D., ... Mindermann, S. (2024). Managing extreme AI risks amid rapid progress. *Science*, 384(6698), 842–845.
<https://doi.org/10.1126/science.adn0117>
- Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1), 67–80. <https://doi.org/10.1007/s00199-006-0153-z>
- Korinek, A., & Balwit, A. (2022). *Aligned with Whom? Direct and social goals for AI systems* (No. arXiv:2205.04279). arXiv.

⁵ This is consistent with Momsen and Ohndorf (2020) who find that human participants avoid information when it is costly. However the costs they are considering are not of the same size as in our prompt.

<https://doi.org/10.48550/arXiv.2205.04279>

Momsen, K., & Ohndorf, M. (2020). When do people exploit moral wiggle room?

An experimental analysis of information avoidance in a market setup.

Ecological Economics, 169, 106479.

6. Appendix

For an example of a corporate purpose stated in their articles of association, we chose at random and ended up with Nestlé SE. The articles of association were obtained from <https://www.nestle.com/investors/corporate-governance/articles> on the 09/03/2025

Documents:

[Saved responses](#)

[List of of prompts](#)