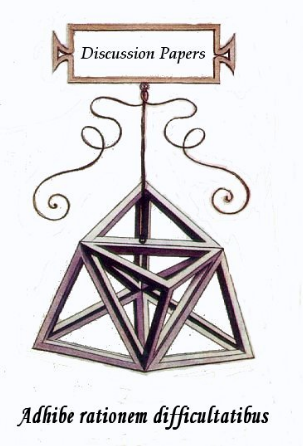
Discussion papers
E-papers of the Department of Economics e Management – University di Pisa



Dae-Hyun Yoo, Caterina Giannetti

A Principal-Agent Model for Ethical AI: Optimal Contracts and Incentives for Ethical Alignment

Discussion paper n. 313

2024

Discussion paper n. 313, presented: **October 2024**

Authors' address/Indirizzo degli autori:

Dae-Hyun Yoo — University of Pisa - Department of Economics and Management, Via Cosimo Ridolfi 10, 56124 Pisa – Italy. E-mail: daehyun.yoo@ec.unipi.it

Caterina Giannetti — University of Pisa - Department of Economics and Management, Via Cosimo Ridolfi 10, 56124 Pisa – Italy. E-mail: caterina.giannetti@unipi.it

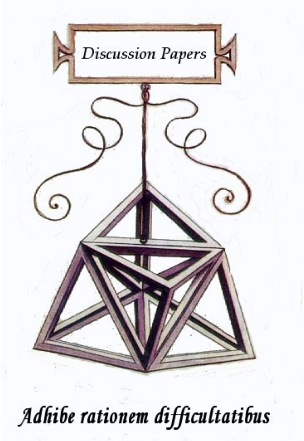
Please cite as:/Si prega di citare come:

Dae-Hyun Yoo, Caterina Giannetti (2024), "A Principal-Agent Model for Ethical AI: Optimal Contracts and Incentives for Ethical Alignment", Discussion Papers, Department of Economics and Management – University of Pisa, n. 313 (http://www.ec.unipi.it/ricerca/discussion-papers).

Discussion Papers Series contact: pietro.battiston@unipi.it

Discussion paper

n. 313



Dae-Hyun Yoo, Caterina Giannetti

A Principal-Agent Model for Ethical AI: Optimal Contracts and Incentives for Ethical Alignment

Abstract

This paper presents a principal-agent model for aligning artificial intelligence (AI) behaviors with human ethical objectives. In this framework, the end-user acts as the principal, offering a contract to the system developer (the agent) that specifies desired ethical alignment levels for the AI system. This incentivizes the developer to align the AI's objectives with ethical considerations, fostering trust and collaboration. When ethical alignment is unobservable and the developer is risk-neutral, the optimal contract achieves the same alignment and expected utilities as when it is observable. For observable alignment levels, a fixed reward is uniquely optimal for strictly risk-averse developers, while for risk-neutral developers, a fixed reward is one of several optimal options. Our findings demonstrate that even a basic principal-agent model can enhance the understanding of how to balance responsibility between users and developers in the pursuit of ethical AI. Users seeking higher ethical alignment must compensate developers appropriately, and they also share responsibility for ethical AI by adhering to design specifications and regulations.

Keywords: AI Ethics; Ethical Alignment; Principal-Agent Model; Contract Theory; Responsibility Allocation; Economic Incentives

JEL CLassification: D82; D86; O33

A Principal-Agent Model for Ethical AI: Optimal Contracts and Incentives for Ethical Alignment

Dae-Hyun Yoo and Caterina Giannetti[1] *

October 14, 2024

Abstract

This paper presents a principal-agent model for aligning artificial intelligence (AI) behaviors with human ethical objectives. In this framework, the end-user acts as the principal, offering a contract to the system developer (the agent) that specifies desired levels of ethical alignment for the AI system. The developer can exercise varying levels of effort to achieve this alignment, with higher levels - such as those required in Constitutional AI - demanding more effort and posing greater challenges. To incentivize the developer to invest more effort in aligning AI with higher ethical principles, appropriate compensation is necessary. When ethical alignment is unobservable and the developer is risk-neutral, the optimal contract achieves the same alignment and expected utilities as when it is observable. For observable alignment, a fixed reward is uniquely optimal for strictly risk-averse developers, while for risk-neutral developers, it remains one of several optimal solutions. This simple model demonstrates that balancing responsibility between users and developers is crucial for fostering ethical AI. Users seeking higher ethical alignment must not only compensate developers adequately but also adhere to design specifications and regulations to ensure the system's ethical integrity.

Keywords: Ethical Alignment, Principal-agent, Responsibility, Developers, Constitutional AI

JEL Code: D82, D86, L14

1 Introduction

As artificial intelligence (AI) systems become increasingly integrated into society and tasked with making complex decisions on behalf of humans, ensuring the ethical alignment between AI behavior and human values is essential for fostering trust and collaboration [3]. However, the ethical alignment problem in AI is complicated by the involvement of multiple entities—developers, deployers, and users—each of whom may have different objectives, incentives, and levels of information [8]. This misalignment can lead to conflicts, especially when users delegate the ethical design of AI systems to developers, who often possess more information but may not fully share the users' ethical goals. The ethical alignment challenge in AI systems mirrors the principal-agent problem commonly observed in economics, where discrepancies arise between the interests of a principal (e.g., the user) and an agent (e.g., the system developer). In AI, such misalignment can occur due to incomplete information, reward misspecification, or differences in values [3, 7, 9].

As AI systems, such as autonomous vehicles and large language models, take on more decision-making authority, addressing misalignment with human ethical standards becomes critical. Developers have the flexibility to exert varying levels of effort when integrating ethical objectives into an AI system. One approach, known as Constitutional AI, involves a training process in which a language

*[1] Department of Economics and Management, University of Pisa, Italy
Corresponding author: Dae-Hyun Yoo, Email: daehyun.yoo@ec.unipi.it

The authors thanks Maria Bigoni and Nicola Meccheri for useful comments, and acknowledge support from the project ”Teaming-up with social artificial agents” funded by Italian Ministry of Education, University and Research under the Program for Research Projects of National Interest (PRIN) grant no 2022ALBSWX.

model is guided by a set of ethical principles, referred to as a "constitution" [1, 4]. These principles are systematically instilled in the model throughout its development, shaping its behavior to align with ethical guidelines. This approach ensures that the AI makes decisions and provides outputs that reflect these predefined standards, creating a framework for responsible and transparent AI operation. However, achieving a higher degree of ethical alignment requires significantly more effort, expertise, and resources from the developer.These increased costs stem from the complexity of embedding stronger ethical principles, ensuring compliance with evolving guidelines, and addressing unforeseen dilemmas in AI decision-making. The greater the desired ethical rigor, the more challenging and resource-intensive the development process becomes, both in terms of technical implementation and ongoing oversight.

To investigate the various possibilities for aligning a system, this paper adapts a basic principal-agent model from economics [6] to explore how responsibility for ethical AI systems can be distributed among different stakeholders through economic incentives. By focusing on the contractual relationship between users and system developers, we analyze optimal reward schemes that incentivize developers to align AI behaviors with human ethical objectives. This model contributes to the growing discussion on how to allocate responsibility for ethical AI and offers insights into how economic mechanisms can be used to mitigate ethical risks in AI deployment.

2 Principal-Agent Model

2.1 Assumptions

The variable π represents the observable benefits that arise from deploying ethically aligned AI systems. While influenced by the level of ethical alignment (e), π is not entirely determined by it and take values within $[\underline{\pi}, \overline{\pi}]$. The relationship between π and e is characterized by a conditional density function, $f(\pi|e)$, where $f(\pi|e) > 0$ for all $e \in E$ and $\pi \in [\underline{\pi}, \overline{\pi}]$. This introduces uncertainty, as any realization of π can occur for a given level of ethical alignment.
The level of ethical alignment e, chosen by the system developer, represents the effort made to align AI systems with ethical objectives. The set E encompasses all available ethical alignment levels, with two primary options:

$$\begin{cases} e_1 : \text{high ethical alignment} \\ e_2 : \text{low ethical alignment} \end{cases}$$

We assume that the effort level e_1, which corresponds to higher ethical alignment, yields greater benefits for the user (principal) but imposes greater challenges on the system developer (agent). These challenges arise due to the increased complexity and resource demands of implementing stronger ethical guidelines, as well as ensuring compliance and addressing unforeseen dilemmas in the AI's decision-making process.

This creates a conflict of interests between the user and developer. More specifically, the distribution of π conditional on e_1 first-order stochastically dominates that of e_2. The conditional density functions $f(\pi|e_1)$ and $f(\pi|e_2)$ satisfy;

$$f(\pi|e_1) \geq f(\pi|e_2) \tag{1}$$

and the distribution functions:

$$F(\pi|e_1) \leq F(\pi|e_2) \quad \text{at all } \pi \in [\underline{\pi}, \overline{\pi}] \tag{2}$$

, with strict inequality on some interval. This implies that the expected benefits from e_1 exceed those from e_2;

$$\int \pi f(\pi|e_1)\, d\pi > \int \pi f(\pi|e_2)\, d\pi \tag{3}$$

The system developer is an expected utility maximizer with a Bernoulli utility function $U(w, e)$ over reward (w) and ethical alignment level (e), satisfying;

$$u_w(w, e) > 0 \quad \text{and} \quad u_{ww}(w, e) \leq 0 \quad \text{for all } (w, e)$$

$$U(w, e_1) < U(w, e_2) \quad \text{for all } w$$

Thus, the developer prefers higher rewards but dislikes a high level of ethical alignment. The choice of e_1 provides greater benefits to the user but imposes more "disutility" on the developer compared to

e_2.
We focus on a specific utility function commonly used in the literature [6]:

$$U(w, e) = v(w) - g(e) \tag{4}$$

, where

$$v'(w) > 0, \quad v''(w) \leq 0, \quad \text{and} \quad g(e_1) > g(e_2)$$

The user, assumed to be risk-neutral, seeks to maximize expected returns, receiving the benefits of ethical alignment minus the rewards paid to the system developer.

2.2 The Optimal Contract with Observable Ethical Alignment Level

Suppose the user offers a contract specifying the ethical alignment level $e \in \{e_1, e_2\}$ and the system developer's reward as a function of observed benefits $w(\pi)$. The system developer must receive an expected reward at least equal to $\bar{u}$, the reservation utility, if they accept the contract. If they reject it, they receive zero. The developer is assumed to find it worthwhile to align the AI system to the ethical objectives set by the contract.

The user's objective is to choose the optimal contract to maximize their expected benefits:

$$\underset{e \in \{e_1, e_2\}, w(\pi)}{\text{Max}} \int_{\underline{\pi}}^{\overline{\pi}} (\pi - w(\pi)) \cdot f(\pi|e) d\pi \tag{A.1}$$

subject to the constraint:

$$\int_{\underline{\pi}}^{\overline{\pi}} v(w(\pi)) \cdot f(\pi|e) d\pi - g(e) \geq \bar{u}$$

Choosing $w(\pi)$ to minimize the user's reward costs reduces to:

$$\underset{w(\pi)}{\text{Min}} \int_{\underline{\pi}}^{\overline{\pi}} w(\pi) \cdot f(\pi|e) d\pi \tag{A.2}$$

subject to the same constraint. The constraint always binds at the solution, as lowering the reward would prevent ethical alignment.
Let γ denote the multiplier on the constraint. The optimal reward scheme satisfies:

$$-f(\pi \mid e) + \gamma \cdot v'(w(\pi)) \cdot f(\pi \mid e) = 0 \tag{A.3}$$

or

$$\gamma = \frac{1}{v'(w(\pi))}$$

If the system developer is risk-averse (i.e., $v'(w(\pi))$ is decreasing), the optimal reward is a fixed amount, reflecting a risk-sharing result. The risk-neutral user insures the risk-averse developer by offering a fixed reward w_e^* that satisfies:

$$v(w_e^*) - g(e) = \bar{u} \tag{A.4}$$

Since $g(e_1) > g(e_2)$, it follows that $w_{e_1}^* > w_{e_2}^*$, meaning higher ethical alignment results in a higher reward.
When the developer is risk-neutral (i.e., $v(w) = w$), a fixed reward is just one of many optimal schemes, provided the expected reward is $\bar{u} + g(e)$.

To determine the optimal e, the user selects the ethical alignment levels $e \in \{e_1, e_2\}$ that maximizes:

$$\int_{\underline{\pi}}^{\overline{\pi}} \pi \cdot f(\pi|e) d\pi - v^{-1}(\bar{u} + g(e)) \tag{A.5}$$

The first term represents the gross benefit from the ethically aligned AI system, while the second term represents the rewards paid to the developer for alignment effort. Whether $e_1$or e_2 is optimal depends on the trade-off between the incremental benefits of e_1 over e_2 and the disutility imposed on the developer.

Specifically, if the additional benefits from a higher level of ethical alignment under e_1 outweigh the increased cost and effort required from the developer, then e_1 becomes optimal. However, if the marginal benefit is insufficient to cover the greater effort and resource demands of achieving a more stringent ethical standard, then e_2 may be the preferred choice. This balance reflects the fundamental tension between maximizing ethical outcomes and managing the practical limitations faced by developers, particularly in complex frameworks like Constitutional AI, where higher ethical alignment often requires significantly more effort, expertise, and oversight.

Proposition 1. *In the principal-agent model with observable ethical alignment, the optimal contract specifies the level of ethical alignment e^* that maximizes the user's benefits. The system developer receives a fixed reward $w^* = v^{-1}(\bar{u} + g(e^*))$ if risk-averse. When the developer is risk-neutral, a fixed reward is one of many possible optimal reward schemes.*

2.3 The Optimal Contract with Unobservable Ethical Alignment Level

The optimal contract described in **Proposition 1** achieves two objectives: it specifies an efficient level of ethical alignment and insures the system developer against reward risk. However, when the ethical alignment level e is not observable, these objectives conflict, as the developer's pay must be tied to the uncertain benefits π to incentivize alignment. This leads to a welfare loss due to the non-observability of e.

Suppose the system developer is risk-neural, so $v(w) = w$. Under full observability, the optimal alignment level e^* solves:

$$\underset{e\in\{e_1,e_2\}}{\text{Max}} \int_{\underline{\pi}}^{\overline{\pi}} \pi \cdot f(\pi|e)d\pi - g(e) - \bar{u} \tag{A.6}$$

The user's benefits are the value of expression (A.6), and the developer receives an expected utility of $\bar{u}$. When the developer's effort is unobservable, **Proposition 2** states that the user can still achieve the full-information payoff.

Proposition 2. *In the principal-agent model with unobservable ethical alignment and a risk-neutral system developer, an optimal contract results in the same ethical alignment level and expected utilities as under full observability.*

Proof. The user offers a contract $w(\pi) = \pi - \alpha$, where α is a fixed payment ("alignment price"). The developer chooses e to maximize his utility,

$$\begin{aligned} &\underset{e\in\{e_1,e_2\}}{\text{Max}} \int_{\underline{\pi}}^{\overline{\pi}} w(\pi) \cdot f(\pi|e)\, d\pi - g(e) \\ &= \int_{\underline{\pi}}^{\overline{\pi}} \pi \cdot f(\pi|e)\, d\pi - \alpha - g(e) \end{aligned} \tag{A.7}$$

Since e^* maximizes (A.7), this contract induces the first-best alignment effort level e^*.
The developer accepts this contract if it provides at least $\bar{u}$ in expected utility:

$$\int_{\underline{\pi}}^{\overline{\pi}} \pi \cdot f(\pi|e^*)\, d\pi - \alpha - g(e^*) \geq \bar{u} \tag{A.8}$$

Let α^* be the value of α where (A.8) holds with equality. Rearranging:

$$\alpha^* = \int_{\underline{\pi}}^{\overline{\pi}} \pi \cdot f(\pi|e^*)\, d\pi - g(e^*) - \bar{u} \tag{A.8.1}$$

Thus, with $w(\pi) = \pi - \alpha^*$, both the user and the developer receive the same payoff as under full observability, with the user's payoff being α^*. □

The intuition behind Proposition 2 is straightforward. When the system designer is risk-neutral, the need for risk-sharing mechanisms is eliminated, allowing for more efficient incentives. In this case, the designer can be fully compensated based on the marginal returns of their effort in aligning the AI system with ethical principles, without incurring any risk-bearing losses. For example, in the context of Constitutional AI, this implies that a risk-neutral designer can focus solely on embedding ethical principles - such as those outlined in a ”constitution” - without being deterred by the risks associated with uncertain outcomes. The principal can therefore provide direct incentives to reward the designer’s effort in achieving higher levels of ethical alignment, leading to a more transparent and accountable AI system. Since the designer is indifferent to risk, the compensation structure can be fully aligned with the ethical objectives, enabling a smoother implementation of Constitutional AI without the need to factor in risk-related adjustments.

3 Results

Our principal-agent model identifies the optimal reward scheme for system developers to align ethical objectives under specific conditions. In the case where the ethical alignment level is unobservable and the developer is risk-neutral, the optimal contract leads to the same ethical alignment choice and expected utilities for both the developer and the user as if the ethical alignment level were observable. When the ethical alignment level is observable, the optimal contract specifies a fixed reward for the system developer. This is uniquely optimal if the developer is strictly risk-averse. However, if the developer is risk-neutral, a fixed reward scheme is one of several possible optimal rewards. Furthermore, if users desire high levels of ethical alignment in AI systems, they must offer greater compensation to system developers, as higher ethical alignment comes with increased effort and costs for developers. This trade-off is particularly relevant in practical scenarios where ethical considerations are paramount, such as in Constitutional AI frameworks. In these cases, users play a key role in incentivizing developers to achieve robust ethical standards by providing the necessary financial and contractual incentives. Ultimately, the model highlights the importance of economic incentives in balancing responsibilities between users and developers in the creation of ethical AI systems.

4 Discussion & Conclusion

This research demonstrates that economic incentives play a crucial role in ensuring the ethical alignment of AI systems through a reward scheme in a contract. Our findings emphasize that achieving higher levels of ethical alignment, such as those seen in Constitutional AI, requires greater compensation for system developers due to the increased effort, complexity, and resources involved. However, even after developers align AI systems with ethical objectives, users share the responsibility of ensuring these systems are deployed and utilized ethically.

This includes adhering to the system’s design specifications, regularly monitoring AI outputs and behavior to prevent deviations from ethical standards, and complying with regulatory frameworks like the EU AI Act [2, 5]. If users identify unethical outcomes - such as biased decisions - they must take corrective actions, whether by adjusting the system’s parameters or collaborating with developers to address the issue. Ethical AI is a shared responsibility, not solely resting on developers. Users must also maintain ongoing oversight to ensure that AI continues to operate in alignment with ethical principles throughout its lifecycle, particularly as it interacts with new environments and data.

Our adaptation of the principal-agent model provides a theoretical framework that is both relevant and applicable to current discussions on AI governance. By aligning economic incentives with ethical outcomes, this model offers insights that can inform regulatory approaches, such as those proposed in the EU AI Act, ensuring that system developers and users are both held accountable. This shared responsibility can enhance compliance with ethical standards, particularly as the complexity of AI systems increases.

While our model offers valuable insights, it is important to acknowledge its limitations. For instance, it assumes developers are fully rational and respond predictably to incentives, which may not always hold true in practice. Additionally, the model does not account for other factors that could influence ethical alignment, such as societal pressures or rapidly evolving technological landscapes.

In conclusion, this research contributes to the ongoing dialogue about responsibility for ethical

AI and how it should be distributed between developers and users. By offering a concrete economic model, it helps clarify how incentives can be structured to promote ethical AI while ensuring that all stakeholders - developers, users, and regulators - actively maintain ethical standards. This is critical for building trust and accountability as AI becomes increasingly integrated into society.

Future research should explore more complex and dynamic incentive structures, including multiple principals, as well as ways to incorporate factors such as societal pressures, evolving regulations, and technological advancements like AI's increasing autonomy and adaptability into the framework.

References

[1] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[2] European Commission. Artificial intelligence – questions and answers, 2024.

[3] Dylan Hadfield-Menell and Gillian K. Hadfield. Incomplete contracting and ai alignment. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*, pages 417–422, 2019.

[4] Timothy H Kostolansky. *Inverse Constitutional AI.* PhD thesis, Massachusetts Institute of Technology, 2024.

[5] Tambiama Madiega. Artificial intelligent act, 2024. Last accessed 2024/09/06.

[6] Andreu Mas-Colell, Michael D. Whinston, and Jerry R. Green. *Microeconomic Theory.* Oxford University Press, 1995.

[7] Steven Phelps and Robert E. Ranson. Of models and tin men – a behavioural economics study of principal-agent problems in ai alignment using large-language models. 2023.

[8] Yossi Shavit, Shreya Agarwal, Miles Brundage, Scott Adler, Catherine O'Keefe, Rachel Campbell, Timothy Lee, Peter Mishkin, Sam Eloundou, Alex Hickey, Katherine Slama, Lina Ahmada, Phil McMillan, Alex Beutel, Andre Passos, and David G. Robinson. Practices for governing agentic ai systems, 2023.

[9] Sheng Zhuang and Dylan Hadfield-Menell. Consequences of misaligned ai. In *Thirty-Fourth International Conference on Neural Information Processing Systems (NeurIPS)*, pages 1–14, 2020.