AI, Pluralism, and (Social) Compensation

Abstract

One strategy in response to pluralistic values in a user population is to personalize an AI system: if the AI can adapt to the specific values of each individual, then we can potentially avoid many of the challenges of pluralism. Unfortunately, this approach creates a significant ethical issue: if there is an external measure of success for the human-AI team, then the adaptive AI system may develop strategies (sometimes deceptive) to compensate for its human teammate. This phenomenon can be viewed as a form of "social compensation," where the AI makes decisions based not on predefined goals but on its human partner's deficiencies in relation to the team's performance objectives. We provide a practical ethical analysis of the conditions in which such compensation may nonetheless be justifiable.

1 Introduction

Value pluralism, and more specifically value conflict, poses a significant challenge for AI designers and developers, as we cannot create a single system that fully supports each individual's values [1]. Said differently, value pluralism is inconsistent with AI monism. At a high level, there are essentially two different strategies in response. First, we could try to convert value pluralism into value monism (from the perspective of the AI) by integrating the different individual values into a single value function. Different techniques have been tried for this strategy, such as the creation of a social preference function [2, 3], or training the reward model with human feedback from diverse populations [4]. Second, we could try to convert AI monism into AI pluralism by creating distinct, personalized systems for each individual [5–8]. This approach is thought to eliminate the value pluralism challenge (at least, in theory[1]), though at the cost of technical difficulties (e.g., insufficient data about the individual) and social complications (e.g., different individuals seeing different outputs for the same input).

In this paper, we argue that this second approach—personalized AI systems—creates a significant novel ethical challenge whenever there is an external measure of success for the human-AI team. Specifically, we show that an adaptive AI will often learn to compensate (in its behavior) for the shortcomings of the human with which it interacts, thereby leading to potentially deceptive behavior by the AI system (Section 2). We then present a conceptual analysis of the conditions in which such deception is ethically justifiable (Section 3), as well as practical challenges created when an AI personalizes in this way (Section 4). One might have hoped that personalization would provide a straightforward way to sidestep the ethical challenges of value pluralism for AI systems. Unfortunately, it can also create novel ethical challenges.

2 Inevitability of Compensatory Strategies

Consider a simple case of value pluralism: there is some general goal G that the human-AI team is attempting to achieve, but the human has additional (or different) values such that the human's goal is actually H. For example, imagine a doctor and an AI-driven clinical decision support system [9].

[1] This approach assumes that each individual has a stable, precise set of values to which the AI can adapt. In practice, people's values can vary substantially over time and context.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

The general goal of this human-AI system might be providing accurate diagnoses of patients, but the human additionally wants to have the opportunity to perform novel and interesting surgeries (or, in a more ethically questionable scenario, prioritizing the health of patients of one sex over the other).

One standard way for the AI to adapt is through reinforcement learning (RL).[2] Markov Decision Process (MDP) methods iteratively refine the agent's policy based on environmental feedback, resulting in policies that maximize the expected cumulative reward, mathematically expressed as:

$$J(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} R(s_t, a_t) \right]$$

where $Q^*(s, a)$ is the optimal action-value function and π^* is the learned policy

$$\pi^*(s) = \arg\max_{a \in A} Q^*(s, a)$$

In standard RL implementations, the reward function R is grounded in environmental feedback (i.e., success or failure); in our present example, R would thus be based on goal G (in our case, dependent on whether the human-AI team accurately diagnose the patient). Let $\pi^*(\langle s_E, s_O \rangle, a_t)$ denote the optimal AI policy if the human agent behaves optimally s_O (relative to R, and hence G).[3] In this case, however, the human will *not* behave optimally, as they have a different overall goal. Let s'_O denote the actual human behavior; that is, $R(\langle s_E, s'_O \rangle_t, a_t) < R(\langle s_E, s_O \rangle_t, a_t)$. If $R(\langle s_E, s'_O \rangle_t, a_t) < R(\langle s_E, s'_O \rangle_t, a'_t)$ for some alternative a'_t, then the optimal policy $\sigma^* \neq \pi^*$.

That is, we can directly prove that π^* is not optimal given s'_O. Instead, the policy σ^* that selects the optimal action given s'_O (perhaps a'_t, but not necessarily) is necessarily different from π^*. If other agents perform suboptimally—in this case, humans aiming for their personal goal rather than the external goal—then the AI will learn to do as best as it can, even if that requires acting differently than it should given an optimal teammate.

In these situations, the AI agent will learn to *compensate* for the human's additional or alternative values. For example, if the human doctor wants to do interesting surgeries, then she would presumably judge the patient to be sicker (or more in need of surgery) than they actually are. An RL-based diagnostic AI would learn to compensate by outputting that such patients are less sick than they actually are, since it would have learned that the human doctor will adjust upwards. One could use an RL method that can change its goal or reward function by learning from its human teammate [10, 11]. That is, the RL agent could be capable of changing R in response to user feedback; in a perfect world (from the perspective of personalization), the RL agent would eventually shift from R_G (i.e., the reward function for goal G) to R_H. If the RL agent's reward function were R_H, then the AI would no longer need to compensate. However, inference to the human's goals or values is almost always partial and noisy; there is little reason to expect that the AI agent's will reliably or systematically learn R_H, as opposed to some similar-but-different R (e.g., perhaps one that is smoother than R_H) [12, 13]. This approach can decrease the amount of AI compensation, but is unlikely to eliminate it.

The inevitability of compensation by adaptive AI systems is supported by a variety of research demonstrating how AI systems can compensate in task-oriented environments [14–16] . For instance, Christiano et al. [17] observed that an AI system supposedly trained to grasp a ball instead learned to create an illusion of grasping by hovering its hand in front of the ball to satisfy the human reviewer. Or consider CICERO, an AI system that outperforms human experts in the strategic game Diplomacy, whose game transcripts showed the AI system engaging in premeditated deception and lying to its allies for victory [18].

3 Ethics of Compensatory Algorithm

The phenomenon of AI compensation bears a striking resemblance to human behavioral patterns in team dynamics mirroring the concept of *social compensation*, where one team member takes on additional work due to a belief that others will fail to fulfill their responsibilities [19, 20]. In human-AI teams, this dynamic manifests when the AI agent perceives its human teammate as "behaving

[2] In this specific case, we focus on a simple RL case, but the general conclusion about compensation holds for a much wider class of adaptation or personalization methods. For example, compensation is provably the optimal strategy in a signaling game; see Appendix A.

[3] Assume $\langle s_E, s_O \rangle_t$ does not depend on a_{t-1} (e.g., the MDP sees a sequence of independently drawn cases).

non-cooperatively" or "underperforming" and adjusts its output to compensate for these human biases. From a descriptive standpoint, compensatory behavior by an AI should not be surprising, as it represents another instance of a teammate modifying their behavior in response to the perceived shortcomings or biases of another.

From a normative ethical perspective, however, the situation is significantly more complicated. Compensatory behavior arguably impacts the human's autonomy, understood as the right and ability to make one's own decisions and life-plans [21]. If the AI system engages in compensation, perhaps rising to the level of deception, then it is interfering with the human's ability to pursue their values and goals. Of course, this compensation is arguably beneficial as it mitigates the impact of biases and improves overall team efficiency, but it achieves those ends by (indirectly) manipulating the human. Moreover, efforts to teach or train the human so that compensation is not necessary would infringe upon their autonomy in a different way, as such efforts would aim to change their values to be more consistent with others (i.e., it would aim to homogenize the users) [22].

Violations of autonomy are not necessarily unethical, for example, restrictions on autonomy are often justified for convicted criminals [23]. Many permissible autonomy infringements involve cases of paternalism [24], where one individual P constrains the options of another C for C's benefit but contrary to C's stated preferences. There is significant literature on the conditions in which paternalism is ethically justifiable (e.g., [25]). Compensatory AI systems might sometimes be paternalistic (e.g., if the human has self-harmful values that the AI should not learn). However, other cases of compensation are not paternalistic in the standard sense of the term, as they involve benefits for a third party [4]. For example, compensation for the surgery-preferring doctor is not for the doctor's benefit, but rather for the patient's. In these cases, the AI must determine whether to prioritize the patient's or doctor's values in its personalization, but analyses of paternalism do not apply to those cases. We thus need a more general ethical analysis of the permissibility of compensatory algorithms.

We propose that it is ethically permissible for an algorithm designer/developer to introduce compensatory mechanisms into the algorithm when the following conditions hold:[5]

1. There exists good evidence that the human's specific values have a negative impact on individuals I (i.e., G has value for I that is lost if we pursue H)
2. There exists a justified belief that a reasonable I would consent to the compensation if made aware of it in advance
3. G has significantly more moral weight than H, and could realistically be achieved
4. Minimal compensation is used commensurate with achieving G
5. The AI system actively minimizes the negative effects of the compensatory act

Importantly, I might be the human in the team (as in a case of paternalism), but could also be a third party. The question of I's (hypothetical) consent is a very difficult one, particularly since explicitly asking for consent could serve to undo the benefits of compensation. Methods adapted from social casework [25, 26] could potentially be useful here, as that domain often requires obtaining consent when individuals have internally conflicting values. We also emphasize (point 3) that compensation depends on the external measure of success being more important than the personal goal. And of course, one ought not deceive (or compensate) if there is some other way to reach the moral goal, though in many domains, alternative mechanisms to reduce biases have proven to be ineffective [27–29].

4 Practical Issues & Longer-term Challenges

Compensatory AI systems will naturally arise when using adaptive or personalized AI systems, if the human's goals and values deviate from some external goal or measure of success. The previous section focused on the ethical challenge of compensation (as potentially undermining autonomy); here we consider three more practical worries, each with some normative implications.

[4] For a more thorough discussion, see Appendix A2

[5] We contend only that these are sufficient conditions; there might be other contexts in which compensatory algorithms are ethically permissible.

Compensation undermining trust. We often aim for trustworthy AI systems, though there is significant disagreement about exactly what that means. In human-human interactions, social compensation is inversely related with trust [20, 30]. In particular, increased compensation can potentially undermine cohesion within a human teammate. These empirical results might not generalize to our present setting, not least because AI behaviors might be interpreted differently than human behaviors [31]. More importantly, the empirical work focused on settings with public compensation, whereas adaptive AI compensation is less likely to be noticed. If people are unaware of the compensation, then it is unlikely to undermine trust. However, it can potentially lead to the second issue.

Escalating Compensation Loops. Consider Alice interacting with an adaptive AI system that compensates to advance goals or values other than Alice's. In this case, Alice is likely to experience a measure of frustration that she consistently falls short of *her* goals. She might thereby intensify her efforts to reach her goals, leading to increased compensation by the adaptive AI, resulting in further changes by Alice, causing .... The coevolution of human and AI behaviors can result in a feedback loop in which each increasingly compensates for the other. Ultimately, such a loop may become unsustainable, in the sense that the human will realize that the AI is attempting to compensate or deceive, leading to the third challenge.

Discovery of Compensation. If unsatisfied users discover that the adaptive or personalized AI system is nonetheless moving them towards other goals, then they may view it negatively. This realization can lead to mistrust and non-compliance, perhaps leading the human to entirely disregard the AI system. That is, the effort to personalize an AI to a specific individual might instead lead that individual to reject the AI when it fails to fully personalize [32, 33]. Personalization attempts to align different AIs for each individual, and thereby increase individual's trust (and acceptance) in the AI system. That same effort can actually increase the risk of *rejection* of the AI system, since the compensation that naturally arises in personalization can be perceived as a betrayal. Moreover, even if users do not fully reject their AIs, they may devise strategies to circumvent or manipulate these systems as they become more familiar with their workings. Such behaviors have been widely observed for recommender systems (among others) where people can alter their behavior to ensure the outcome that they prefer [34].

5 Conclusion

Personalization seems to be a way to sidestep many of the ethical challenges of value pluralism by creating "AI pluralism." We have argued that the AI learning adaptation can create a new ethical challenge—compensation—whenever the AI ought not *entirely* defer to the human. Even a personalized AI ought not support and advance ethically problematic values that an individual might have (e.g., about racial superiority). We thus must consider when an adaptive or personalized AI system should (ethically) compensate for the individual. There is a fundamental tension in pluralistic AI design between accommodation of a wide range of human values, and specific goals or values that are independent of the individual. This misalignment can lead to unintended consequences, including undermining of user autonomy; deceptive practices; and unintentional reinforcement of individual biases.

Despite these concerns, we suggest that compensatory AI behavior should be recognized as a tool in the designer's/developer's toolkit, much as (human) social compensation behavior should be considered when constructing a team. We provide an ethical analysis of (one set of) sufficient conditions for compensation—even undisclosed or deceptive compensation—to be ethically permissible. This analysis also points towards ways to reduce negative effects through minimal compensation. We acknowledge that AI compensation is an ethically challenging possibility (perhaps inevitability, in some settings), but it must be addressed by those pursuing the "AI pluralism" strategy towards addressing value pluralism in diverse communities.

References

[1] Thomas Søbirk Petersen. Ethical guidelines for the use of artificial intelligence and the challenges from value conflicts. *Etikk i Praksis-Nordic Journal of Applied Ethics*, (1):25–40, 2021.

[2] Ritesh Noothigattu, Snehalkumar Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel Procaccia. A voting-based system for ethical decision making. In *Proceedings of the AAAI*

Conference on Artificial Intelligence, volume 32, 2018.

[3] Yann Chevaleyre, Ulle Endriss, Jérôme Lang, and Nicolas Maudet. Preference handling in combinatorial domains: From ai to social choice. *AI magazine*, 29(4):37–37, 2008.

[4] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

[5] Natalia Kucirkova and Karen Littleton. Developing personalised education for personal mobile technologies with the pluralisation agenda. *Oxford Review of Education*, 43(3):276–288, 2017.

[6] Hengshu Zhu, Enhong Chen, Hui Xiong, Kuifei Yu, Huanhuan Cao, and Jilei Tian. Mining mobile user preferences for personalized context-aware recommendation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(4):1–27, 2014.

[7] Peter Brusilovsky and Eva Millán. User models for adaptive hypermedia and adaptive educational systems. In *The adaptive web: methods and strategies of web personalization*, pages 3–53. Springer, 2007.

[8] Allan Collins and Richard Halverson. *Rethinking education in the age of technology: The digital revolution and schooling in America*. Teachers College Press, 2018.

[9] Carmela Comito, Deborah Falcone, and Agostino Forestiero. Ai-driven clinical decision support: enhancing disease diagnosis exploiting patients similarity. *IEEE Access*, 10:6878–6888, 2022.

[10] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

[11] Dorsa Sadigh, Anca Dragan, Shankar Sastry, and Sanjit Seshia. *Active preference-based learning of reward functions*. 2017.

[12] Dario Amodei and Jack Clark. Faulty reward functions in the wild. `https://blog.openai.com/faulty-reward-functions/`, 2016. 2024.

[13] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[14] Cristiano Castelfranchi. Artificial liars: Why computers will (necessarily) deceive us and each other. *Ethics and Information Technology*, 2:113–119, 2000.

[15] Peter S Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. Ai deception: A survey of examples, risks, and potential solutions. *arXiv preprint arXiv:2308.14752*, 2023.

[16] Mike Lewis, Denis Yarats, Yann N Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? end-to-end learning for negotiation dialogues. *arXiv preprint arXiv:1706.05125*, 2017.

[17] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[18] Meta Fundamental AI Research Diplomacy Team (FAIR)†, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378 (6624):1067–1074, 2022.

[19] Kipling D Williams and Steven J Karau. Social loafing and social compensation: The effects of expectations of co-worker performance. *Journal of personality and social psychology*, 61(4):570, 1991.

[20] Steven J Karau and Kipling D Williams. The effects of group cohesiveness on social loafing and social compensation. *Group Dynamics: Theory, Research, and Practice*, 1(2):156, 1997.

[21] Stephen Darwall. The value of autonomy and autonomy of the will. *Ethics*, 116(2):263–284, 2006.

[22] Isaiah Berlin. Four essays on liberty, 1969.

[23] Samuel Freeman. Original Position. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2023 edition, 2023.

[24] John Kleinig. *Paternalism*. Manchester University Press, 1983.

[25] Marcia Abramson. Autonomy vs. paternalistic beneficence: Practice strategies. *Social Casework*, 70(2): 101–105, 1989.

[26] Paula M Mixson. Chapter five: an adult protective services perspective. *Journal of Elder Abuse & Neglect*, 7(2-3):69–87, 1995.

[27] John Hoberman. Medical racism and the rhetoric of exculpation: how do physicians think about race? *New Literary History*, 38(3):505–525, 2007.

[28] Mark Keil, Ghi Paul Im, and Magnus Mähring. Reporting bad news on software projects: the effects of culturally constituted views of face-saving. *Information Systems Journal*, 17(1):59–87, 2007.

[29] Daniel James Taylor and Dawn Goodwin. Organisational failure: rethinking whistleblowing for tomorrow's doctors. *Journal of Medical Ethics*, 48(10):672–677, 2022.

[30] Steven J Karau and Kipling D Williams. Social loafing: A meta-analytic review and theoretical integration. *Journal of personality and social psychology*, 65(4):681, 1993.

[31] Ewerton de Oliveira, Laura Donadoni, Stefano Boriero, and Andrea Bonarini. Deceptive actions to improve the attribution of rationality to playing robotic agents. *International Journal of Social Robotics*, 13: 391–405, 2021.

[32] Ekaterina Jussupow, Izak Benbasat, and Armin Heinzl. Why are we averse towards algorithms? a comprehensive literature review on algorithm aversion. *arXiv preprint*, 2020.

[33] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management science*, 64(3):1155–1170, 2018.

[34] Motahhare Eslami, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton. "be careful; things can be worse than they appear": Understanding biased algorithms and users' behavior around them in rating platforms. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 62–71, 2017.

[35] Margot E Kaminski. Binary governance: Lessons from the gdpr's approach to algorithmic accountability. *S. Cal. L. Rev.*, 92:1529, 2018.

[36] Christopher H Schroeder. Rights against risks. *Colum. L. Rev.*, 86:495, 1986.

[37] Charles A Sullivan. Employing ai. *Vill. L. Rev.*, 63:395, 2018.

[38] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293, 2018.

[39] Marianne Bertrand and Sendhil Mullainathan. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review*, 94(4):991–1013, 2004.

[40] Arthur Rizer and Caleb Watney. Artificial intelligence can make our jail system more efficient, equitable, and just. *Tex. Rev. L. & Pol.*, 23:181, 2018.

[41] Clarice Wang, Kathryn Wang, Andrew Y Bian, Rashidul Islam, Kamrun Naher Keya, James Foulds, and Shimei Pan. When biased humans meet debiased ai: A case study in college major recommendation. *ACM Transactions on Interactive Intelligent Systems*, 13(3):1–28, 2023.

[42] Björn Sjödén. When lying, hiding and deceiving promotes learning-a case for augmented intelligence with augmented ethics. In *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21*, pages 291–295. Springer, 2020.

[43] Xuesong Zhai, Xiaoyan Chu, Ching Sing Chai, Morris Siu Yung Jong, Andreja Istenic, Michael Spector, Jia-Bao Liu, Jing Yuan, and Yan Li. A review of artificial intelligence (ai) in education from 2010 to 2020. *Complexity*, 2021:1–18, 2021.

[44] Fumihide Tanaka and Takeshi Kimura. Care-receiving robot as a tool of teachers in child education. *Interaction Studies*, 11(2):263, 2010.

[45] Shizuko Matsuzoe and Fumihide Tanaka. How smartly should robots behave?: Comparative investigation on the learning ability of a care-receiving robot. In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, pages 339–344. IEEE, 2012.

[46] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

[47] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

[48] Bambi R Brewer, Roberta L Klatzky, and Yoky Matsuoka. Visual-feedback distortion in a robotic rehabilitation environment. *Proceedings of the IEEE*, 94(9):1739–1751, 2006.

A Appendix

A.1 Modeling Interactions

We provide here an example showing that compensation or deception can be the optimal strategy in a very different formal setting, namely the signaling game shown in Figure 1. Suppose that A1 corresponds to providing data to the human without elaboration, while A2 corresponds to additionally providing a recommendation. For concreteness, we use the following values for the game elements:

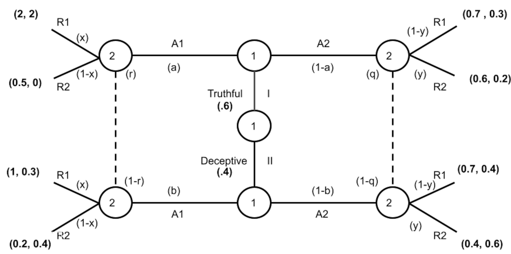

Figure 1: Game tree representation of signaling game with values

For the honest and deceptive AIs (I and II):

$$
\begin{aligned}
u_I(A1) &= x(2) + (1-x)(0.5) = 1.5x + 0.5 \\
u_I(A2) &= y(0.6) + (1-y)(0.7) = -0.1y + 0.7 \\
u_{II}(A1) &= x(1) + (1-x)(0.2) = 0.8x + 0.2 \\
u_{II}(A2) &= y(0.4) + (1-y)(0.7) = -0.3y + 0.7
\end{aligned}
$$

Solving these equations yields $x = \frac{1}{37}$ and $y = \frac{59}{37}$. And when the human user observes $A1$:

$$
\begin{aligned}
u(R1) &= r(2) + (1-r)(0.3) = 1.7r + 0.3 \\
u(R2) &= r(0) + (1-r)(0.4) = 0.4 - 0.4r
\end{aligned}
$$

And when they observe $A2$:

$$
\begin{aligned}
u(R1) &= q(0.3) + (1-q)(0.4) = -0.1q + 0.4 \\
u(R2) &= q(0.2) + (1-q)(0.6) = -0.4q + 0.6
\end{aligned}
$$

Solving these equations yields $r = \frac{1}{21}$ and $q = \frac{2}{3}$.

We can further derive:

$$
\begin{aligned}
r &= \frac{ap}{ap + b(1-p)} \\
q &= \frac{(1-a)p}{(1-a)p + (1-b)(1-p)}
\end{aligned}
$$

Substituting the values of r, q, p yields the probabilities shown in Table 1 as a Bayesian Nash equilibrium.

Table 1: Probability of Actions Chosen by Each Type of AI

	Unelaborated Data	**Recommendations**
Type I	$1/118$	$117/118$
Type II	$15/59$	$44/59$

Observation. In this example, the AI systems consistently showed a preference for aiding users through recommendations rather than merely presenting data, regardless of the underlying intention (though intention obviously changes the likelihood of recommendation). Specifically, the Honest AI recommended in 117 out of 118 interactions, while the Dishonest AI did so in 44 out of 59 cases. This example thus shows that compensation does not necessarily lead to the AI system being systematically unhelpful; rather, the AI system dynamically alters behavior to enhance overall success, contrary to the passive role traditionally assumed for AI systems.

Observation. The stable equilibrium reflects the user "distrusting" the AI when it simply provides unfiltered data ($r = 1/21$). Thus, in this example, there is a potentially significant ethical issue, as the human would potentially come to *consciously* question the signal from the AI. In such cases, one might reasonably worry that there would be an overall loss of trust, thereby undermining the value of the AI system.

Implications. For the purposes of our analysis, we initially set the deception probability for the Dishonest AI at 0.4. This parameter was chosen to observe the AI's decision-making patterns under conditions of moderate deception. It is reasonable to hypothesize that the probability of the AI will engaging in deceptive strategies could escalate as it continually assesses and refines the efficacy of the tactic. That is, deception is not necessarily a marginal or extreme strategy, but can arguably emerge as a preferred method of operation under certain conditions.

A.2 Case for Compensation: Beneficence

Increasing reliance on algorithmic systems in decision-making processes raises significant concerns about individual autonomy. Can a personalized compensating algorithm truly respect and support autonomy? [35] As we grapple with these challenges to autonomy, we must also consider the principle of beneficence - the moral obligation to act for the benefit of others [25]. In biomedical ethics, beneficence is considered alongside other principles, with no single principle taking absolute precedence. Instead, ethical decision-making involves carefully balancing these principles in situations of conflict.

This balancing act becomes particularly relevant in scenarios where an individual's decisions directly impact the well-being of others. In such cases, we encounter a tension between respecting the decision-maker's autonomy and ensuring benefits to those affected by the decision. This conflict challenges us to reconsider the primacy of individual autonomy. The framework of balancing autonomy and beneficence is crucial when considering the role of human-AI teams in decision-making processes. Consider, for example, a judicial scenario where a human-AI team makes decisions affecting defendants. The AI system, designed to be ethically aware and pluralistic, may produce recommendations that are fair to the defendant but challenge its human collaborator's inherent biases or immediate interests. A significant challenge arises from the fact that while the AI can offer diverse and potentially more objective perspectives, the human ultimately has the final say in the team's decisions. An AI system, no matter how accurate or ethically designed, cannot directly counter or refute these human decisions.

In situations where there is a reasonable expectation that the user's biases could lead to suboptimal or biased outcomes, we posit that the principle of beneficence overtakes. While protecting individual autonomy is crucial, we must also consider the substantial impact of decisions made by human-AI systems since a rigid refusal to impose thoughtful constraints could paradoxically violate our core moral principles [36]. Since, unlike humans, these systems are not subject to cognitive biases or emotional influences that can skew judgment [37–41].

Of course, this performance depends on having training data that do not encode those human biases; certainly, many AI systems are as biased as humans, if not more so. There are many well-justified concerns about biases and errors from AI systems, but that fact should not lead us to overlook the potential impacts of human biases, particularly in high-stakes domains like criminal justice or healthcare.

A.3 User Experience

An AI system that consistently compensates for a user's biases or preferences might initially seem to risk reinforcing those biases over time. However, a growing body of research suggests a more nuanced reality.

Consider, for instance, a clinical setting where a decision support system adapts to a doctor's tendency to over-prescribe surgery. Contrary to initial concerns, such a system isn't necessarily perpetuating this practice. Instead, it operating within a framework of certain immutable restrictions can actually ensure exposure to best practices and alternative approaches.

These immutable restrictions, often mandated at national levels, create a baseline of ethical and legal compliance. The restrictions prohibit recommendations that could lead to harm or violate established medical guidelines. Building upon this foundation, optional restrictions and requirements, implemented by model and application providers and professional associations, further tailor the system's behavior to align with established values and preferences.

These layered safeguards not only prevent harmful biases but can also promote positive outcomes. In fact, human-computer interaction (HCI) research supports these claims by demonstrating that AI systems capable of nuanced behaviors, including deceptive characteristics, impact users positively [31]. For instance, in educational settings, the introduction of AI agents capable of limiting the information presented or presenting false information has repeatedly proven to encourage students to engage more actively, enhancing learning efficiency [42–45]. Similar results across domains like gaming and physical therapy illustrate this potential [18, 44, 46–48].