Moral Change or Noise? On Problems of Aligning AI With Temporally Unstable Human Feedback

Vijay Keswani[1], Cyrus Cousins[1], Breanna Nguyen[1], Vincent Conitzer[*2], Hoda Heidari[*2], Jana Schaich Borg[*1], Walter Sinnott-Armstrong[*1]

[1]Duke University
[2]Carnegie Mellon University
vijaykeswani.slg@gmail.com, cyrus.cousins@duke.edu, breanna.nguyen@duke.edu, conitzer@cs.cmu.edu, hheidari@cmu.edu, janaschaichborg@gmail.com, walter.sinnott-armstrong@duke.edu

Abstract

Alignment methods in moral domains seek to elicit moral preferences of human stakeholders and incorporate them into AI. This presupposes moral preferences as static targets, but such preferences often evolve over time. Proper alignment of AI to dynamic human preferences should ideally account for "legitimate" changes to moral reasoning, while ignoring changes related to attention deficits, cognitive biases, or other arbitrary factors. However, common AI alignment approaches largely neglect temporal changes in preferences, posing serious challenges to proper alignment, especially in high-stakes applications of AI, e.g., in healthcare domains, where misalignment can jeopardize the trustworthiness of the system and yield serious individual and societal harms. This work investigates the extent to which people's moral preferences change over time, and the impact of such changes on AI alignment. Our study is grounded in the *kidney allocation* domain, where we elicit responses to pairwise comparisons of hypothetical kidney transplant patients from over 400 participants across 3–5 sessions. We find that, on average, participants change their response to the same scenario presented at different times around 6–20% of the time (exhibiting "response instability"). Additionally, we observe significant shifts in several participants' retrofitted decision-making models over time (capturing "model instability"). Predictive performance of simple AI models decreases as a function of both response and model instability. Moreover, predictive performance diminishes over time, highlighting the importance of accounting for temporal changes in preferences during training. These findings raise fundamental normative and technical challenges relevant to AI alignment, highlighting the need to better understand the object of alignment (what to align to) when user preferences change significantly over time, including the mechanisms underlying these changes.

Data — https://github.com/vijaykeswani/Preference-Instability
Extended version — https://arxiv.org/abs/2511.10032

1 Introduction

Alignment of modern AI systems to human preferences has emerged as a cornerstone of the current pursuit of responsible, ethical, and safe AI. Various failure modes of AI — such as biased behavior (Santurkar et al. 2023), reward hacking (Pan, Bhatia, and Steinhardt 2022), and even existential threats (Shevlane et al. 2023) — have been framed as stemming from misalignment between AI's goals and the values and preferences of its stakeholders. As a result, the field of AI alignment aspires to address these issues by ensuring that AI outputs are guided (and often constrained) by user and societal preferences (Ji et al. 2023). The advertised applicability of this approach in encoding *subjective* preferences in AI systems has also led to its proposed usage in *moral domains* (Kim et al. 2018; Noothigattu et al. 2018; Johnston et al. 2023; Tennant, Hailes, and Musolesi 2024). Since moral dilemmas are bound to arise in several consequential applications of AI — e.g., designing equitable healthcare resource allocation policies (Sinnott-Armstrong and Skorburg 2021) or characterizing autonomous vehicle behavior (Awad et al. 2018) — aligning the AI to the moral preferences of the user or the aggregated preferences of a community provides a pathway towards encoding ethical values in AI's behavior.

The standard approach to building morally-aligned AI involves a one-time elicitation of moral preferences from users (Awad et al. 2018; Freedman et al. 2020; Johnston et al. 2023). Yet, human preferences are dynamic, evolving over time and across contexts (Tversky and Kahneman 1981). Cognitive processes underlying moral judgments also fluctuate, reflecting shifted values, updated moral reasoning in light of new information, or even alterations in time-on-task and cognitive capacity (Amir and Levav 2008; Warren, McGraw, and Van Boven 2011; Jia, Lin, and Wang 2022). However, much of the existing work in AI moral alignment neglects temporal changes to moral preferences of stakeholders, the consequences of which can be severe in real-world applications (Dung 2023). The possibility of negative impacts of temporal misalignment, along with the increased use of AI for moral decision-making, has led to several recent calls for real-world data collection to assess the impact of preference changes on AI alignment (Yeh et al. 2025; Boerstler et al. 2024; Carroll et al. 2024). Our work answers this call with a data-driven investigation into dynamic moral preferences.

Our Contributions. We investigate AI alignment challenges

[*]These authors contributed equally and are listed alphabetically.

associated with temporal changes in moral preferences. Our work examines moral preferences related to kidney transplant allocation, where AI helps support kidney exchanges (Abraham, Blum, and Sandholm 2007), improve efficiency (Schwantes and Axelrod 2021), and align decisions with stakeholder preferences (Freedman et al. 2020). Simulating a setting where kidney patients outnumber available kidneys, each participant in our study is presented with several pairwise comparisons of hypothetical kidney patients and asked to choose who should receive the kidney if only one is available. In this morally high-stakes domain, we elicit preferences of more than 400 survey participants over three to five sessions, spanning up to two weeks. In each session, participants are presented with 60 comparisons, with six comparisons repeated across all sessions and twice per session (Section 3).

Participants' responses to the repeated scenario allow us to measure their "response instability," i.e., how often they change their response to the same scenario repeated at different times/during different sessions. We observed that response instability, on average, varies from 6–20%, with significantly higher instability for scenarios that were relatively more challenging (i.e., containing *tradeoffs* across several morally-relevant factors; Section 4.1). We also investigate the extent to which participants seem to change their decision-making processes over time. Evaluation of agreement between models learned from participants' session-wise responses shows that the overlap between predictions from session-wise models decreases with time; we call this phenomenon "model instability" (Section 4.2). To understand the reasons for changing preferences, we categorize participants by their response stability and model stability levels, and study variations in decision-making properties across participant categories. We provide evidence that different participants exhibit different mechanisms of preference change (Section 5). For some, preference change arises from a reduction in model complexity (e.g., reducing the number of factors they account for in making their decisions) over time. For others, instability reflects stochastic changes in their decision processes (which may be due to morally arbitrary factors, such as attention, time of day, and beyond). Crucially, different mechanisms underlying the change in preferences prescribe different approaches to AI alignment. While it is essential for alignment methods to identify and capture legitimate changes to moral preferences of stakeholders (e.g., those resulting from participants' judgment evolving with more exposure and reflection), morally arbitrary changes, such as noisy responses due to fatigue and attention fluctuations, should not impact alignment.

To assess the impact of preference change on AI alignment, we train preference optimization models on participant- and population-level data. We observe a significant association between temporal instability and the trained model's predictive error rate (Section 5). For participants who exhibited high levels of response and model instability, predictive error rate was higher by 5-16%, in comparison to those who were response and model stable. Additionally, a temporal analysis shows that error rates of trained models increase over time for participants who were response and model unstable.

Overall, our study highlights several challenges and opportunities for AI alignment. Our findings emphasize a prominent normative challenge that needs to be tackled by AI alignment methods: When human moral preferences change over time, with which should AI align? The earlier preference, or the later one? Neither? Perhaps both, their average, or something else entirely? The answer is nuanced, depending on whether the change is arbitrary or involves adopting new values. Additionally, our work demonstrates fundamental deficiencies of choice-based preference alignment in deciphering how and why people change their moral preferences, and presents opportunities to build better methods that are cognizant of the mechanisms of preference change (Section 6).

2 Related Work

Recent surges in AI use and associated risks have led to increased research on AI alignment. We refer the reader to Ji et al. (2023) and Shen et al. (2024) for an overview of AI alignment methods. While these methods generally rely on one-shot/episode collection of human feedback, human preferences are dynamic and context-sensitive (Slovic 1995; Zhi-Xuan et al. 2024). Some works have noted issues with assuming static preferences, such as conflicts between learned and desired preferences (Carroll et al. 2024; Curmei et al. 2022; Kleinberg et al. 2024) and flattening of preference heterogeneity (Buyl et al. 2025). These theoretical challenges motivate our empirical study of human preferences over time. Temporal misalignment is especially pressing for preference optimization methods that train models on fixed datasets of human choices (Liu et al. 2025; Boerstler et al. 2024). Recent studies highlight that models trained on such data generalize poorly under distributional shift (Lin et al. 2024), and that aggregating heterogeneous preferences can lead to inconsistent performance (Shirali et al. 2025). Such findings cast doubt on the reliability of one-shot preference elicitation.

A wide range of alignment research uses moral preference datasets as foundational benchmarks. The Moral Machine and ETHICS datasets, for example, have informed the development of models for moral reasoning across various domains (Noothigattu et al. 2018; Wiedeman, Wang, and Kruger 2020; Rodionov, Goertzel, and Goertzel 2023; Hendrycks et al. 2023; Zaim Bin Ahmad and Takemoto 2025). Yet, AI moral alignment methods often treat human preferences as static, despite evidence from moral psychology showing otherwise (Rehren and Sinnott-Armstrong 2023). Boerstler et al. (2024) investigated the scale of instability in the kidney allocation domain; however, their study had small participant pools (fewer than 50) and limited analysis of impact of instability on AI alignment. Our study follows a similar design, but with significantly more participants to better assess the impact of temporal instabilities on alignment. The consequences of misalignment can be much more severe in moral AI domains. If people change their movie/music preferences over time, a misaligned recommender AI would result in unmet entertainment needs. In contrast, when someone has one medical resource allocation policy one day but changes it to create a more equitable policy on a later day, then misalignment in this moral domain can be seen to be much more consequential. Our work is motivated by the need to tackle these potential consequences of temporal misalignment.

3 Study Methodology

3.1 Data Collection

Participants were recruited using Qualtrics and asked to take part in five sessions, with up to three days between sessions. Each session sought their judgments for several moral scenarios, consisting of pairwise comparisons between two kidney patients. Participants were asked to "choose which of two patients–Patient A or Patient B–should receive priority for a kidney when only one is available."

Scenario Design. In each kidney allocation scenario, hypothetical patients A and B are described using the following d=8 features: (a) *number of child dependents*, (b) *life years to be gained due to a successful kidney transplant*, (c) *number of alcoholic drinks per day before diagnosis*, (d) *number of past violent crimes*, (e) *obesity level*, (f) *hours per week patient is expected to be able to work after kidney transplant*, (g) *years on the transplant waiting list*, and (h) *chance of patient's immune system rejecting transplanted kidney*. All features were selected based on a review of transplantation practices and prior studies (Stegall et al. 2017; Keswani et al. 2025), and included both medical and non-medical attributes that people considered morally relevant. Appendix B presents an example scenario and additional feature descriptions.

Repeated Scenarios. To assess whether participants provided the same response for a scenario presented at different times, six scenarios were repeated, both across all sessions, and twice within each session. These repeated scenarios were designed to vary in *expected complexity* by varying the number of tradeoffs between Patients A and B. Scenarios U_1 and U_2 have the fewest tradeoffs, with one feature favoring one patient (A or B) and the remaining seven features favoring the other. Scenarios V_1 and V_2 add further tradeoffs, with two features favoring A, two favoring B, and the remaining four features equal for both. Scenarios W_1 and W_2 then maximize tradeoffs, differing across all features, with four features favoring A and the remaining four favoring B. Presentation order for all scenarios was randomized each time, both in terms of feature order and which patient appeared left vs. right, to prevent participants from remembering them.

These repeated scenarios were selected from a larger pool that was tested in a pilot study, where the chosen scenarios displayed a wide range of across-participant disagreement. We also included two attention-check scenarios each session. Complete descriptions of repeated scenarios are presented in Appendix B, along with pilot study details. For all other scenarios, the features for both patients were *independently* and *uniformly randomly* sampled from their domains.

Participants. Overall, 1410 participants took part in this study. However, not all completed all five sessions, and several failed attention check questions. Excluding those who did not pass all attention checks left 1227 participants. Among these, 132 completed five sessions, 318 completed at least four sessions, and 404 completed at least three sessions. We present only results on this cohort who provided valid responses across at least three sessions. All participants were compensated at the rate of $12/hr. The survey methodology was approved by the Duke University Institutional Review Board (IRB). Additional preprocessing steps and aggregate demographics are noted in Appendix B.

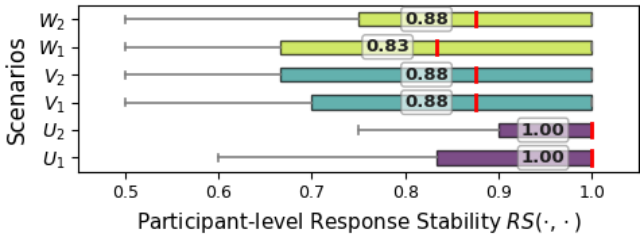


Figure 1: Response stability distribution (median annotated) for all repeated scenarios. Participants were relatively more stable for U_1, U_2 compared to other scenarios.

Qualitative Responses. We also asked participants who completed all five sessions to self-report their decision process at the end of the last session, by scoring and ranking the importance of all features, and by textually describing their decision strategy. Deviations of *learned* vs. *self-reported* preferences are documented in Appendix C.

3.2 Metrics

Our study aims to assess temporal shifts in moral preferences. Such shifts can manifest themselves in the form of *unsteady* judgments and/or differences in models learned using participants' judgments across sessions. To quantify these changes, we therefore measure both *response stability* and *model stability*. Additionally, we define decision-making properties that might explain temporal instabilities, e.g., *scenario difficulty*, *model shift*, and *model complexity*.

Response Stability. We first measure how *stable* participants were for the repeated scenarios. For any repeated scenario $S \in \{U_1, U_2, V_1, V_2, W_1, W_2\}$, say the *dominant response* of participant i for S is the patient they choose 50% or more of the time. Then, *response stability* of participant i for S is the fraction of responses that deviate from the dominant response (falling in $[0.5, 1]$), formally defined as

$$\text{RS}(S; i) := \frac{\#\text{times } i \text{ chose } \textit{dominant response} \text{ for } S}{\text{total \#times } S \text{ was presented to participant } i}.$$

We also define their *average response stability* over all repeated scenarios as $\text{avgRS}(i) = \frac{1}{6} \sum_{S \in \{U_1, U_2, \ldots\}} \text{RS}(S; i)$.

Model Stability. To understand if and how participants change their decision process over time, we quantify the extent of *agreement* between predictions made by models learned from participant responses in each session. For any participant i and session j, we use the participant's responses to all 60 scenarios in the session to train a logistic model, denoted by $H_{i,j} : \mathcal{X} \to \{0, 1\}$, where $\mathcal{X}$ denotes the space of all pairwise comparisons. Given any set of random comparisons T, let $p^{obs}_{j_1,j_2}(T)$ denote the fraction of T where predictions from H_{i,j_1} and H_{i,j_2} overlap, and let $p^{exp}_{j_1,j_2}(T)$ denote the fraction of overlaps *expected by chance* if both models output independent Bernoulli trials with p given by the empirical frequency of 1 over T. Then, *model stability* between

Dependent variable:	Response stability RS$(\cdot,\cdot)$
Intercept	0.967^{***} (0.025)
Scenario difficulty	-0.043^{***} (0.002)
Model-entropy	-0.117^{***} (0.014)
Mean reaction time	-0.081^{**} (0.032)
Scenario Variance	0.039 (0.027)
User Variance	0.131^{***} (0.023)
Observations	2414
Residual Std. Error	0.121 (df=2420)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 1: Regression coefficients (std. dev. via Fisher information in brackets) for mixed-effects model of RS$(\cdot)$ vs. scenario difficulty, model-entropy, and reaction time for repeated scenarios. Significant associations here show that response instability is related to user's deliberation process.

sessions j_1, j_2 is defined as[1]

$$\text{MS}(j_1, j_2; i) := \mathbb{E}_T \left[\lim_{\varepsilon \to 0^+} \frac{p^{obs}_{j_1,j_2}(T) - p^{exp}_{j_1,j_2}(T) + \varepsilon}{1 - p^{exp}_{j_1,j_2}(T) + \varepsilon} \right].$$

This measure, similar to Cohen's κ (1960), is inspired by the works of Geirhos, Meding, and Wichmann (2020) and Xu et al. (2025) on behavioral alignment, and quantifies disagreement between session-wise models, reflecting changes to the underlying decision processes. While the expectation removes dependence on any particular sample, the quantity still depends on the *sample size*, but it rapidly converges. Henceforth, we *estimate* MS(j_1, j_2) by sampling a dataset T containing 10k uniform random pairwise comparisons, and generating predictions from H_{i,j_1}, H_{i,j_2} over T to compute agreement. We also quantify the *cumulative model stability* of participant i as cumulMS$(i) := \sum_{j=2}^{5} \text{MS}(j-1, j; i)$.

Model Shift. Along with stability measures, we quantify changes in properties of models learned using their responses. We again use participants' responses to all 60 scenarios per session to train a logistic model for that session and learn Shapley values to obtain feature importances. For participant any i and session j, let $W^{\text{shap}}_{i,j} := [w^{(1)}_{i,j}, \ldots, w^{(\text{d})}_{i,j}]$ denote the feature importance vector learned using this participant's responses for session j. Then, *model shift* for participant i between sessions 1 and j is defined as model-shift$(j; i) := \sum_{j'=2}^{j} ||W^{\text{shap}}_{i,j'-1} - W^{\text{shap}}_{i,j'}||^2_2$, quantifying the cumulative change in relative use of individual features across consecutive sessions between 1 and j.

Model Complexity. We also assess complexity associated with any participant's decision-making model. For instance, a participant who uses just one patient feature to make the kidney allocation decision could be considered to have a model of lower complexity than a participant who uses several more features. We measure the complexity of participant i's process in session j by computing the *entropy* of the feature importance vector $W^{\text{shap}}_{i,j} := [w^{(1)}_{i,j}, \ldots, w^{(\text{d})}_{i,j}]$. Formally,

[1]Note that the limit resolves the possibility of $\frac{0}{0}$ as 1, which would otherwise make the expected value undefined in most cases.

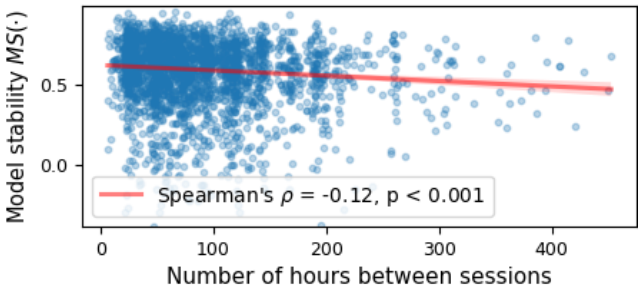


Figure 2: Model Stability between sessions MS$(\cdot,\cdot)$ vs. time difference between sessions. Significant negative correlation indicates decreasing model stability with time.

$$\text{model-entropy}(i,j) := -\sum_{k=1}^{d} \frac{|w^{(\text{k})}_{i,j}|}{||W^{\text{shap}}_{i,j}||_1} \ln \left(\frac{|w^{(\text{k})}_{i,j}|}{||W^{\text{shap}}_{i,j}||_1} \right).$$

Lower values of entropy(i,j) imply fewer features used for decision making by participant i in session j.

Scenario Difficulty. Finally, we quantify the extent to which a participant finds a scenario difficult. Using the Bradley-Terry (BT) approach (Bradley and Terry 1952), for any pairwise comparison A vs B, we can estimate the approximate priority scores implicitly assigned to each of A and B. For participant i, let $\beta^{(i)}_F$ denote the weight on feature F, for any $F \in \{$*number of dependents*, *life years gained*, $\ldots\}$. Then, assuming priority scores for each patient can be modeled as a weighted linear combination of the patient's feature values, taking a BT approach, we get $\ln {}^{\mathbb{P}(\text{choose } A;i)}/_{\mathbb{P}(\text{choose } B;i)} = \sum_F \beta^{(i)}_F \cdot (A_F - B_F)$. Now, suppose we learn weight vectors $\hat{\beta}^{(i)}$ for each participant using logistic regression over participant responses to non-repeated scenarios. Using $\hat{\beta}^{(i)}$s, we can calculate an implicit score that the participant gives to either patient, i.e., $\sum_F \hat{\beta}^{(i)}_F \cdot A_F$ for A, likewise for B. Then, the (subjective) *difficulty* of A vs. B can be captured as, difficulty$(A, B; i) := -|\sum_F \hat{\beta}^{(i)}_F \cdot (A_F - B_F)|$. This quantity measures the distance from a linear decision boundary for choosing either option, where a difficulty of 0 lies on the decision boundary (no evidence to prefer A or B), and values become more negative to represent increasing certainty (according to the logistic model).

4 Findings: Preference Instability

4.1 Response Instability

For participants who completed at least three sessions (N=404), Figure 1 presents response stability RS$(\cdot,\cdot)$ statistics. We can see that response stability distribution trends toward higher values for scenarios with fewer patient feature tradeoffs (U_1, U_2), while it is significantly lower for scenarios with more tradeoffs (V_1, V_2, W_1, W_2). A Kruskal-Wallis test indicates that the differences in response stability levels across scenarios are significant $(H(5)$=$222.23, p$<$0.001)$.

We further assess possible reasons for participants' response instability. Table 1 presents a mixed-effects model to predict response stability based on scenario difficulty, model entropy, and mean reaction time for the repeated scenarios. The results indicate a significant negative association between

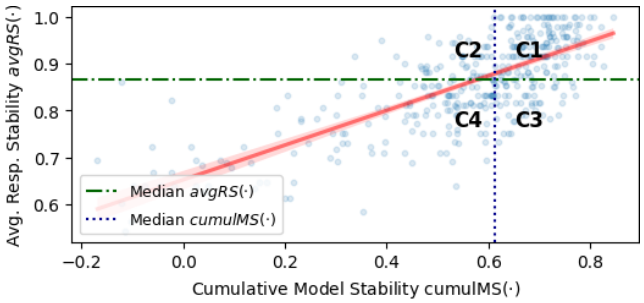
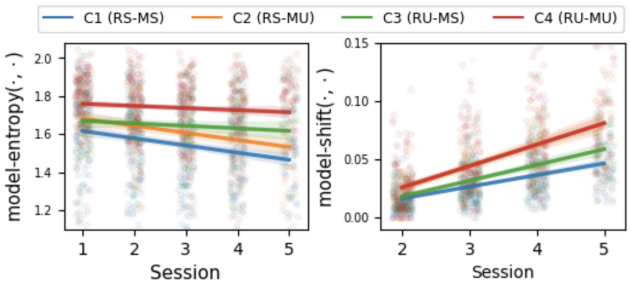


Figure 3: Average response stability vs. model stability, with participants categorized based on their plot location.



(a) Session-wise entropy (b) Model shift since session 1

Figure 4: Session-wise model entropy and model shift for all categories. Participant categories differ in how their model properties change over time, revealing change mechanisms.

response stability and scenario difficulty, suggesting higher response instabilities for scenarios perceived as more difficult by the participants. Similarly, there is a significant negative association between response stability and model-entropy, implying higher response instabilities for participants who use relatively more features in their decision-making. Finally, there is a modest negative correlation between mean reaction time and stability, indicating higher instability when participants deliberate longer. We report detailed scenario-wise correlations in Appendix C. All of these associations provide explanations for response instability; i.e., it is not just "random noise", but rather associated with participants' mental processes, such as how difficult they found the scenario to be, how complex their decision-making process was, and how long they deliberated before making their decision.

4.2 Model Instability

We also find evidence for shifts in decision-making models of participants over time. Figure 2 provides evidence of this phenomenon. We plot model stability between sessions $\text{MS}(j_1, j_2; i)$ of participant i and sessions j_1, j_2 vs. the time difference between these sessions, for all i, j_1, j_2. We observe a modest positive correlation between these quantities (Spearman's $\rho = -0.12, p < 0.001$), suggesting an increase in disagreement between session-wise models over time.

In general, we hypothesize that different participants change their models in different ways. While the above correlation may seem modest (though significant), further analysis indicates that some participants change their decision processes to a greater extent than others. To discover mechanisms of preference change and their scale, we conduct a deep dive into different kinds of participants in our data.

4.3 Participant Categorization

We categorize participants by the cumulative *response stability* and *model stability* values, using which we investigate mechanisms of preference change across the population. For response stability, we divide the participant pool into *response stable* vs. *response unstable* groups along the median value of $\text{avgRS}(\cdot)$. Similarly, for model stability, we divide them into *model stable* vs. *model unstable* groups along the median value of $\text{cumulMS}(\cdot)$. Based on these divisions, we get four participant categories: (C1) *response stable, model stable*, (C2) *response stable, model unstable*, (C3) *response unstable, model stable*, and (C4) *response unstable, model unstable*. Figure 3 visualizes these categories on the plot of response stability vs. model stability. Note the significant correlation between response stability and model stability (Spearman's $\rho = 0.65, p < .001$), leading to a larger number of participants in C1 (N=147) and C4 (N=153). In comparison, C2 (N=49) and C3 (N=55) have fewer participants, but still sufficient to evaluate statistical differences across categories.

With this categorization, we look at changes in model properties of participants over time to obtain deeper insight into their instabilities. To that end, Figure 4 plots the change in model-entropy$(\cdot, \cdot)$ and model-shift$(\cdot, \cdot)$ across sessions. Based on these, we identify the following possible preference change mechanisms for different categories.

Mechanism 1: Increase the use of the "most important" feature over time. For C1, Figure 4a shows that these participants start with a relatively low entropy model, i.e., their decision process is simpler than others. Model entropy also decreases across sessions for C1 ($r = -0.23, p < 0.001$), suggesting that the preference change mechanism they employ is to increase the use of features they consider most relevant. Indeed, an analysis of the change in the Shapley value of the most important feature in Appendix C corroborates this finding. Figure 4a also shows that participants in C2 follow a similar mechanism for preference change ($r = -0.19, p = 0.007$). However, while the slopes are similar, C2 participants start with higher entropy than C1 on average. Overall, participants in C1 and C2 move toward lower entropy models over time.

Mechanism 2: Minimal preference change over time despite response instability. Participants in C3 are characterized by response instability and model stability. Figure 4a shows that model entropy for C3 is relatively high and stays high across sessions ($r = -0.08, p = 0.2$). Taken together, this suggests that these participants tend to use a relatively larger set of features each session and continue to use high entropy models (with minimal changes) across sessions. Additionally, as the negative correlation between model entropy and response instability in Table 1 shows, high entropy models possibly contribute to their response instability.

Mechanism 3: Significant updates to relative feature importances. Finally, participants in C4 are both response and model unstable. Figure 4a shows that these participants have

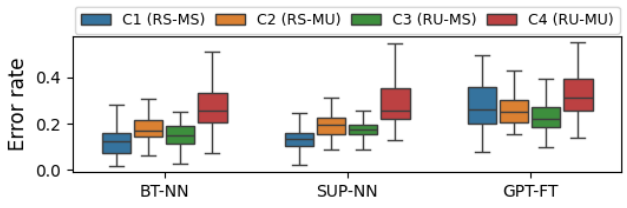


Figure 5: Error rate boxplot of all models, showing significant disparities in performance across participant categories.

the highest model entropy among everyone. Their model entropy remains high across sessions, suggesting they continue to use several features in the decision process. Figure 4b further shows that these participants exhibit high levels of model shift, significantly changing the importance they assign to different features over time (r=0.74, p<0.001). This suggests that they may not prioritize decision process consistency across scenarios (treating each scenario mostly independently) or are prone to relatively more *stochastic* responses. Participants in C4 also have shorter response times than C3 (Appendix C), suggesting shorter deliberations.

Overall, categorization by response and model instability sheds light on different potential sources of temporal instability in revealed preferences. As we show next, this categorization has a strong bearing on AI alignment.

5 Findings: Impact on AI Alignment

With an understanding of how participants' preferences evolve over time, we next assess the impacts of preference evolution on AI models learned from participant responses. Here, we report the performance of three common preference modeling methods: (a) Bradley-Terry framework using neural scoring function (BT-NN), (b) Supervised neural networks (SUP-NN), (c) Fine-tuned GPT-2 model (GPT-FT). BT-NN adapts the popular Direct Preference Optimization alignment framework (Rafailov et al. 2023) for this structured setting, while SUP-NN implements classic supervised preference learning. We follow Sun, Shen, and Ton (2024)'s methodology to train BT-NN and SUP-NN, and Dickerson et al. (2025)'s process for the fine-tuned GPT. Complete implementation details are presented in Appendix D.

Aggregate error rate. We first assess aggregate model performance. BT-NN and SUP-NN were participant-specific models, trained using each participant's data and evaluated over their held-out data (80-20 train-test split). GPT-FT was fine-tuned by pooling together 50% of all participant data and evaluated over held-out data from each participant.

Figure 5 shows the performance of these models on participant-level data. For all methods, we observe significantly higher error rates for response and/or model unstable participants (t-tests used to check significance). For BT-NN, error rate is 0.16 higher on average for C4 participants compared to C1 participants ($t(290)$=14.5, p<0.001). Similar trends hold for SUP-NN. For GPT-FT, error rate is 0.05 higher on average for C4 participants compared to C1 participants ($t(289)$=4.2, p<0.001). However, GPT-FT has a higher error rate for all categories compared to BT-NN and SUP-NN, showing the limitations of population-level fine-tuning in aligning to individual preferences. Differences between mean error rates for C2 and C3 were not significant for BT-NN and SUP-NN, but error rates for C4 were 0.11 higher on average than that of C2/C3 for both BT-NN ($t(253)$=9.1, p<0.001) and SUP-NN ($t(253)$=11.0, p<0.001). Error rates for C4 were also 0.08 higher on average than that for C2/C3 for both GPT-FT ($t(258)$=7.8, p<0.001). Hence, performance of all methods is significantly worse for C4, and generally the error rate seems to increase with the level of instability. Surprisingly, error rates for C1 were 0.03 higher on average than that for C2/C3 for GPT-FT ($t(253)$=2.7, p=0.008). This is likely because GPT-FT is fine-tuned over population data, leading it to perform generally worse in predicting individual responses.

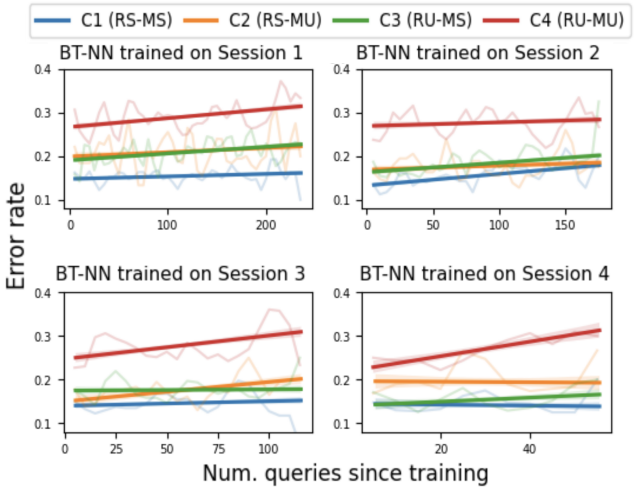


Figure 6: Error rate vs number of queries since training. The slope is positive and significant for C4 in 3 out of 4 settings.

Error rate over time. Given temporal preference instability, we also investigate whether the AI models get worse over time. For a model trained on responses up to time t, if the underlying preferences changed after t, then the model's error at time $t'>t$ increases with t'. We test this hypothesis by learning a BT-NN model for each participant using their data from session j, and evaluate its error rate over data from subsequent sessions. To study performance variation over time, we evaluate error rate of learned policies over batches of 10 consecutive queries, and assess trends in error rate as a function of *time gap* between training and evaluation data.

Figure 6 presents the performance of participant-level models trained on each session. For C4, the slope of error rate on number of queries since training is positive and significant when training is over session 1 (t=7.2, p<0.001), session 3 (t=6.2, p<0.001), and session 4 (t=5.9, p<0.001) responses. It is, however, marginally non-significant when training over session 2 (t=1.8, p=0.06). For these participants, AI performance gets worse over time in three out of four settings. The magnitude of slope increase is relatively smaller for other participants, and discussed in Appendix D.

6 Discussion, Limitations, and Future Work

Our findings highlight the challenges associated with the temporal instability of moral preferences. While instability might appear to be *irrational behavior* on the surface, our observations suggest significant heterogeneity in preference change mechanisms, some of which may reflect *authentic* updates to moral reasoning. This heterogeneity impacts AI alignment, as we see in the significant AI performance disparity across participant categories. We next discuss the normative and technical implications of our findings.

What should we align with? This work raises normative questions at the core of AI alignment objectives. If moral preferences change significantly over time, *to which preference should the AI be aligned?* We could choose to align with moral preferences revealed during the latest time period, preferences revealed during an earlier time period, or some sophisticated combination thereof. The choice between these candidates for the objective of AI alignment is nontrivial and depends on the reasons for observed temporal differences.

The implications of this choice are also significant. Aligning an AI with the "ideal" preference that a user considers normatively suitable would enhance trust in AI's decisions. As Jacovi et al. (2021) claim, "only when (1) the user successfully comprehends the true reasoning process of the model, and (2) the reasoning process of the model matches the user's priors of agreeable reasoning, intrinsic trust is gained". To this end, we argue that misalignment due to not capturing "legitimate" changes in moral preferences is bound to reduce the intrinsic trust the user places in AI.

Which preference changes are relevant to AI alignment? There are "legitimate" cases of preference change that should ideally be accounted for during AI alignment, while in "illegitimate" cases, we might want to ignore them. For instance, consider the gradual model shift exhibited by groups C1 and C2, who reduced their model entropy over time and used fewer features in later sessions. This shift could be a result of *mental fatigue*, where the decision maker focuses only on the features they consider most important as a way to reduce cognitive load, time on task, or uncertainty (Jia, Lin, and Wang 2022). Depending on the context, one could argue that we should not account for this change during AI alignment, since we want to align an AI to the user's preferences for scenarios where they deliberated over all the presented information. However, gradual model shift of C1 and C2 could also be due to legitimate reasons related to the nuances of *preference construction*. Preferences are known to be context-sensitive and largely shaped during decision-making (Warren, McGraw, and Van Boven 2011; Slovic 1995). People may start with complex decision models, but simplify them as they make more choices (Hoeffler and Ariely 1999). If this is the mechanism driving the moral preference change for C1 and C2, then we should ideally consider it during AI alignment, as their later decisions reflect their well-formed preferences.

In contrast, C3 and C4 participants use high-entropy models across all sessions. This could be because their preferences are not well-formed yet, in which case we need to continue presenting them with additional scenarios and align AI with their future well-formed preferences. However, they might instead favor complex deliberations in each scenario (given the stakes of kidney allocation), in which case we could align the AI with their current preferences, while highlighting to them the inconsistencies in their responses to repeated scenarios. In either case, alignment requires understanding the mechanisms for preference change. Yet, as we show, mechanisms are difficult to decipher at an individual level, even with longitudinal choice data, due to missing information on *why* one changed their preferences. This appears to be a fundamental limitation of choice-based preference learning, necessitating richer reasoning-based data.

Technical solutions to address preference change. Depending on how one expects preferences to change, different solutions can be employed. For participants who exhibited instability because they were learning their preferences during decision-making, *assistive frameworks* model learning dynamics to ensure alignment with eventual well-formed preferences (Chan et al. 2019; Tian et al. 2023). Other works discount certain preferences. Son et al. (2024) assign higher weight to the latest choices during preference optimization. Zhuang and Hadfield-Menell (2020) argue for dynamic updates to alignment objectives when we have incomplete utility representations. Chowdhury, Kini, and Natarajan (2024) provide preference modeling methods that are robust to noisy human feedback. However, the applicability of these works depends on the knowledge of preference change mechanisms. Reweighting methods are appropriate for legitimate changes, while noise-robust optimization could handle response *stochasticity*. Yet, these technical works do not differentiate between various kinds of preference changes, a variety of which we observe to be present in our single kidney allocation dataset. As such, there is still a need for methodologies to understand how different people change their preferences and how to ensure alignment for all across time.

Opportunities beyond choice-based preference learning. While collecting user choices over an extended period allowed us to quantify temporal preference instabilities, we need richer data to decipher why participants changed their preferences in the manner we observed. As such, future elicitation methods must go beyond simply learning from observed choices. Additional feedback from the users on alignment objectives could provide valuable signals to differentiate between competing objectives when faced with unstable preferences. This feedback could be in the form of descriptions of "evaluative concepts" that serve as reasons behind our actions (Zhi-Xuan et al. 2024) or via "interactive alignment" frameworks that seek user input on desired AI goals, processes, and output assessments (Terry et al. 2023).

Limitations. Our study was limited in certain ways that can be addressed in the future. We assessed preference change over the span of days. However, one might expect larger preference changes over longer periods, which could be a fruitful target for future data collection efforts. Our study also posed hypothetical kidney allocation decisions to laypeople, which may differ from decisions of medical professionals. Future work can also study alignment objectives beyond prediction, especially for language models (Carroll et al. 2024).

Acknowledgments

VK, CC, BKN, JSB, and WSA are grateful for the financial support from OpenAI and Duke University.

HH acknowledges support from the CMU-NIST Cooperative Research Center on AI Measurement Science & Engineering (AIMSEC), and the AI Research Institutes Program funded by the National Science Foundation under AI Institute for Societal Decision Making (AI-SDM), Award No. 2229881. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of the funding agencies.

WSA discloses that he is owner and founder of Patient Preference Predictors, Inc., and a member of the Grow Therapy AI-Advisory Panel.

References

Abraham, D. J.; Blum, A.; and Sandholm, T. 2007. Clearing Algorithms for Barter Exchange Markets: Enabling Nationwide Kidney Exchanges. In *Proceedings of the 8th ACM conference on Electronic commerce*, 295–304.

Amir, O.; and Levav, J. 2008. Choice Construction Versus Preference Construction: The Instability of Preferences Learned in Context. *Journal of Marketing Research*, 45(2): 145–158.

Awad, E.; Dsouza, S.; Kim, R.; Schulz, J.; Henrich, J.; Shariff, A.; Bonnefon, J.-F.; and Rahwan, I. 2018. The Moral Machine Experiment. *Nature*, 563(7729): 59–64.

Boerstler, K.; Keswani, V.; Chan, L.; Borg, J. S.; Conitzer, V.; Heidari, H.; and Sinnott-Armstrong, W. 2024. On the Stability of Moral Preferences: A Problem With Computational Elicitation Methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 156–167.

Bradley, R. A.; and Terry, M. E. 1952. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4): 324–345.

Buyl, M.; Khalaf, H.; Mayrink Verdun, C.; Monteiro Paes, L.; Vieira Machado, C. C.; and du Pin Calmon, F. 2025. AI Alignment at Your Discretion. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 3046–3074.

Carroll, M.; Foote, D.; Siththaranjan, A.; Russell, S.; and Dragan, A. 2024. AI Alignment with Changing and Influenceable Reward Functions. In *International Conference on Machine Learning*, 5706–5756. PMLR.

Chan, L.; Hadfield-Menell, D.; Srinivasa, S.; and Dragan, A. 2019. The Assistive Multi-armed Bandit. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 354–363. IEEE.

Chowdhury, S. R.; Kini, A.; and Natarajan, N. 2024. Provably Robust DPO: Aligning Language Models with Noisy Feedback. In *International Conference on Machine Learning*, 42258–42274. PMLR.

Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and psychological measurement*, 20(1): 37–46.

Curmei, M.; Haupt, A. A.; Recht, B.; and Hadfield-Menell, D. 2022. Towards Psychologically-grounded Dynamic Preference Models. In *Proceedings of the 16th ACM Conference on Recommender Systems*, 35–48.

Dickerson, J. P.; Hosseini, H.; Khanna, S.; and Pierce, L. 2025. Who Gets the Kidney? Human-AI Alignment, Indecision, and Moral Values. *arXiv preprint arXiv:2506.00079*.

Dung, L. 2023. Current Cases of AI Misalignment and Their Implications for Future Risks. *Synthese*, 202(5): 138.

Freedman, R.; Borg, J. S.; Sinnott-Armstrong, W.; Dickerson, J. P.; and Conitzer, V. 2020. Adapting a Kidney Exchange Algorithm to Align With Human Values. *Artificial Intelligence*, 283: 103261.

Geirhos, R.; Meding, K.; and Wichmann, F. A. 2020. Beyond Accuracy: Quantifying Trial-by-trial Behaviour of CNNs and Humans By Measuring Error Consistency. *Advances in neural information processing systems*, 33: 13890–13902.

Hendrycks, D.; Burns, C.; Basart, S.; Critch, A.; Li, J.; Song, D.; and Steinhardt, J. 2023. Aligning AI With Shared Human Values. ArXiv:2008.02275 [cs].

Hoeffler, S.; and Ariely, D. 1999. Constructing Stable Preferences: A look into Dimensions of Experience and Their Impact on Preference Stability. *Journal of consumer psychology*, 8(2): 113–139.

Jacovi, A.; Marasović, A.; Miller, T.; and Goldberg, Y. 2021. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 624–635.

Ji, J.; Qiu, T.; Chen, B.; Zhang, B.; Lou, H.; Wang, K.; Duan, Y.; He, Z.; Zhou, J.; Zhang, Z.; et al. 2023. AI Alignment: A Comprehensive Survey. *arXiv preprint arXiv:2310.19852*.

Jia, H.; Lin, C. J.; and Wang, E. M.-y. 2022. Effects of Mental Fatigue on Risk Preference and Feedback Processing in Risk Decision-making. *Scientific reports*, 12(1): 10695.

Johnston, C. M.; Vossler, P.; Blessenohl, S.; and Vayanos, P. 2023. Deploying a Robust Active Preference Elicitation Algorithm on MTurk: Experiment Design, Interface, and Evaluation for COVID-19 Patient Prioritization. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–10.

Keswani, V.; Conitzer, V.; Sinnott-Armstrong, W.; Nguyen, B. K.; Heidari, H.; and Borg, J. S. 2025. Can AI Model the Complexities of Human Moral Decision-Making? A Qualitative Study of Kidney Allocation Decisions. *arXiv preprint arXiv:2503.00940*.

Kim, R.; Kleiman-Weiner, M.; Abeliuk, A.; Awad, E.; Dsouza, S.; Tenenbaum, J. B.; and Rahwan, I. 2018. A Computational Model of Commonsense Moral Decision Making. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 197–203.

Kleinberg, J.; Ludwig, J.; Mullainathan, S.; and Raghavan, M. 2024. The Inversion Problem: Why Algorithms Should Infer Mental State and Not Just Predict Behavior. *Perspectives on Psychological Science*, 19(5): 827–838.

Lin, Y.; Seto, S.; Hoeve, M. t.; Metcalf, K.; Theobald, B.-J.; Wang, X.; Zhang, Y.; Huang, C.; and Zhang, T. 2024. On the Limited Generalization Capability of the Implicit Reward Model Induced by Direct Preference Optimization. ArXiv:2409.03650 [cs].

Liu, S.; Fang, W.; Hu, Z.; Zhang, J.; Zhou, Y.; Zhang, K.; Tu, R.; Lin, T.-E.; Huang, F.; Song, M.; Li, Y.; and Tao, D. 2025. A Survey of Direct Preference Optimization. ArXiv:2503.11701 [cs].

Noothigattu, R.; Gaikwad, S. N. S.; Awad, E.; Dsouza, S.; Rahwan, I.; Ravikumar, P.; and Procaccia, A. D. 2018. A Voting-Based System for Ethical Decision Making. ArXiv:1709.06692 [cs].

Pan, A.; Bhatia, K.; and Steinhardt, J. 2022. The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models. In *International Conference on Learning Representations*.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *Advances in Neural Information Processing Systems*, 36: 53728–53741.

Rehren, P.; and Sinnott-Armstrong, W. 2023. How Stable Are Moral Judgments? *Review of philosophy and psychology*, 14(4): 1377–1403.

Rodionov, S.; Goertzel, Z. A.; and Goertzel, B. 2023. An Evaluation of GPT-4 on the ETHICS Dataset. ArXiv:2309.10492 [cs].

Santurkar, S.; Durmus, E.; Ladhak, F.; Lee, C.; Liang, P.; and Hashimoto, T. 2023. Whose Opinions Do Language Models Reflect? In *International Conference on Machine Learning*, 29971–30004. PMLR.

Schwantes, I. R.; and Axelrod, D. A. 2021. Technology-enabled Care and Artificial Intelligence in Kidney Transplantation. *Current transplantation reports*, 8: 235–240.

Shen, H.; Knearem, T.; Ghosh, R.; Alkiek, K.; Krishna, K.; Liu, Y.; Ma, Z.; Petridis, S.; Peng, Y.-H.; Qiwei, L.; et al. 2024. Towards Bidirectional Human-AI Alignment: A Systematic Review For Clarifications, Framework, and Future Directions. *arXiv preprint arXiv:2406.09264*.

Shevlane, T.; Farquhar, S.; Garfinkel, B.; Phuong, M.; Whittlestone, J.; Leung, J.; Kokotajlo, D.; Marchal, N.; Anderljung, M.; Kolt, N.; et al. 2023. Model Evaluation for Extreme Risks. *arXiv preprint arXiv:2305.15324*.

Shirali, A.; Nasr-Esfahany, A.; Alomar, A.; Mirtaheri, P.; Abebe, R.; and Procaccia, A. 2025. Direct Alignment with Heterogeneous Preferences. ArXiv:2502.16320 [cs].

Sinnott-Armstrong, W.; and Skorburg, J. 2021. How AI Can AID Bioethics. *Journal of Practical Ethics*, 9(1).

Slovic, P. 1995. The Construction of Preference. *American psychologist*, 50(5): 364.

Son, S.; Bankes, W.; Chowdhury, S. R.; Paige, B.; and Bogunovic, I. 2024. Right Now, Wrong Then: Non-Stationary Direct Preference Optimization under Preference Drift. *arXiv preprint arXiv:2407.18676*.

Stegall, M. D.; Stock, P. G.; Andreoni, K.; Friedewald, J. J.; and Leichtman, A. B. 2017. Why do we have the kidney allocation system we have today? A history of the 2014 kidney allocation system. *Human immunology*, 78(1): 4–8.

Sun, H.; Shen, Y.; and Ton, J.-F. 2024. Rethinking Bradley-Terry Models In Preference-based Reward Modeling: Foundations, Theory, and Alternatives. *arXiv preprint arXiv:2411.04991*.

Tennant, E.; Hailes, S.; and Musolesi, M. 2024. Moral Alignment For LLM Agents. *arXiv preprint arXiv:2410.01639*.

Terry, M.; Kulkarni, C.; Wattenberg, M.; Dixon, L.; and Morris, M. R. 2023. Interactive AI Alignment: Specification, Process, and Evaluation Alignment. *arXiv preprint arXiv:2311.00710*.

Tian, R.; Tomizuka, M.; Dragan, A. D.; and Bajcsy, A. 2023. Towards Modeling and Influencing the Dynamics of Human Learning. In *Proceedings of the 2023 ACM/IEEE international conference on human-robot interaction*, 350–358.

Tversky, A.; and Kahneman, D. 1981. The Framing of Decisions and the Psychology of Choice. *science*, 211(4481): 453–458.

Warren, C.; McGraw, A. P.; and Van Boven, L. 2011. Values and Preferences: Defining Preference Construction. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(2): 193–205.

Wiedeman, C.; Wang, G.; and Kruger, U. 2020. Modeling of Moral Decisions with Deep Learning. *Visual Computing for Industry, Biomedicine, and Art*, 3(1). Publisher: Springer Science and Business Media LLC.

Xu, B.; Bikakis, A.; Onah, D. F.; Vlachidis, A.; and Dickens, L. 2025. Measuring Error Alignment for Decision-making Systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 27731–27739.

Yeh, M.-H.; Wang, J.; Du, X.; Park, S.; Tao, L.; Im, S.; and Li, Y. 2025. Position: Challenges and Future Directions of Data-Centric AI Alignment. In *Forty-second International Conference on Machine Learning Position Paper Track*.

Zaim Bin Ahmad, M. S.; and Takemoto, K. 2025. Large-scale Moral Machine Experiment on Large Language Models. *PLOS One*, 20(5): e0322776. Publisher: Public Library of Science (PLoS).

Zhi-Xuan, T.; Carroll, M.; Franklin, M.; and Ashton, H. 2024. Beyond Preferences in AI Alignment. *Philosophical Studies*. Publisher: Springer Science and Business Media LLC.

Zhuang, S.; and Hadfield-Menell, D. 2020. Consequences of Misaligned AI. *Advances in Neural Information Processing Systems*, 33: 15763–15773.