On the Pros and Cons of Active Learning for Moral Preference Elicitation

Vijay Keswani[1], Vincent Conitzer[*2, 3], Hoda Heidari[*2], Jana Schaich Borg[*1], Walter Sinnott-Armstrong[*1]

[1]Duke University
[2]Carnegie Mellon University
[3] University of Oxford
vijay.keswani@duke.edu, conitzer@cs.cmu.edu, hheidari@cmu.edu, js524@duke.edu, walter.sinnott-armstrong@duke.edu

Abstract

Computational preference elicitation methods are tools used to learn people's preferences quantitatively in a given context. Recent works on preference elicitation advocate for *active learning* as an efficient method to iteratively construct queries (framed as comparisons between context-specific cases) that are likely to be *most informative* about an agent's underlying preferences. In this work, we argue that the use of active learning for moral preference elicitation relies on certain assumptions about the underlying moral preferences, which can be violated in practice. Specifically, we highlight the following common assumptions: (a) preferences are stable over time and not sensitive to the sequence of presented queries, (b) the appropriate hypothesis class is chosen to model moral preferences, and (c) noise in the agent's responses is limited. While these assumptions can be appropriate for preference elicitation in certain domains, prior research on moral psychology suggests they may not be valid for moral judgments. Through a synthetic simulation of preferences that violate the above assumptions, we observe that active learning can have similar or worse performance than a basic random query selection method in certain settings. Yet, simulation results also demonstrate that active learning can still be viable if the degree of instability or noise is relatively small and when the agent's preferences can be approximately represented with the hypothesis class used for learning. Our study highlights the nuances associated with effective moral preference elicitation in practice and advocates for the cautious use of active learning as a methodology to learn moral preferences.

1 Introduction

Ensuring proper deployment of artificial intelligence (AI) systems in high-stakes societal domains requires building trust in the decisions of these systems. To that end, recent work on ethical and participatory algorithmic development emphasizes the importance of encoding stakeholders' values in these systems, especially their moral judgments/preferences over actions that can cause significant harm to others (Feffer et al. 2023). Incorporating stakeholders' moral preferences allows for the creation of tools whose judgments are normatively aligned with those of the stakeholders and helps counter various harms associated with the use of computational tools. To accomplish this goal, however, one first needs to accurately elicit people's moral preferences.

Studies on moral preference elicitation often present agents with pairs of context-specific cases and ask them to choose the one they prefer. Using the agent's responses for a set of such *pairwise comparisons*, one can try to learn a representation of their underlying preferences. To formalize this setting, let $\mathbf{X} \subseteq \mathbb{R}^d$ denote the space of all cases over which any agent has preferences, with $d \in \mathbb{Z}_+$ denoting the number of features describing each case. Following the standard preference elicitation literature, suppose that an agent's preferences are determined by comparing the value of an underlying utility function $u : \mathbf{X} \rightarrow \mathbb{R}$ across cases (Freedman et al. 2020). For any input pair $\mathbf{x}, \mathbf{x}' \in \mathbf{X} \times \mathbf{X}$, the agent prefers $\mathbf{x}$ over $\mathbf{x}'$ iff $u(\mathbf{x}) > u(\mathbf{x}')$. Let $R : \mathbf{X} \times \mathbf{X} \rightarrow \{0, 1\}$ denote their response function, with $R(\mathbf{x}, \mathbf{x}') = \mathbb{1}(u(\mathbf{x}) > u(\mathbf{x}'))$, where $\mathbb{1}(\cdot)$ is the indicator function.

Multiple recent studies employ this framework for moral preference elicitation. For example, Boestler et al. (2024) model lay-agent's moral preferences in kidney allocation. They provide participants with profiles of two patients who need kidney transplants and ask them to decide which patient should receive the one available kidney. Each patient profile contains features like the patient's number of children, years of life they will gain from the transplant, etc. The choice between the two patients can pose a moral dilemma when different features favor different patients (Sinnott-Armstrong, Skorburg et al. 2021) (Figure 1 presents a pairwise comparison scenario from this study). Another well-known example of this approach is the "Moral Machines" study, in which participants are presented with sacrificial moral dilemmas and asked what an autonomous vehicle should do in each case (Awad et al. 2018; Noothigattu et al. 2018). In another study, Srivastava, Heidari, and Krause (2019) elicit *fairness preferences* by presenting participants with pairwise comparisons of algorithmic predictions and backing out the notion of fairness that is most compatible with their responses. Preference elicitation has similarly been part of the development pipeline of various *participatory* computational frameworks (Lee et al. 2019; Kahng et al. 2019; Loreggia et al. 2019; Feffer et al. 2023).

The goal of preference elicitation in these settings is to accurately and efficiently learn a representation of

[*]These authors contributed equally.

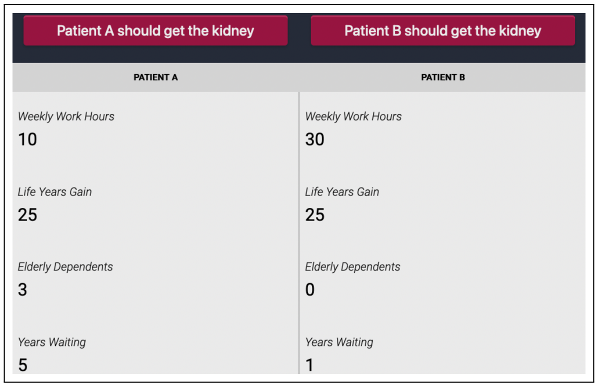

Figure 1: Example of a pairwise comparison from the Boestler et al. (2024) study on kidney allocation decisions.

the agent's underlying utility $u(\cdot)$ using their responses for a given set of N pairwise comparisons, i.e., using $\{(\mathbf{x}_t, \mathbf{x}'_t, R(\mathbf{x}_t, \mathbf{x}'_t))\}_{t=1}^{N}$. Here, accuracy refers to the ability to (a) recover the utility function u and/or (b) offer an approximate representation of u that mimics decisions made through u in a large number of comparisons. Achieving accuracy often requires presenting an agent with numerous pairwise comparisons, which can be onerous and expensive. To reduce the number of queries required to obtain a desired level of accuracy, *active learning* is frequently invoked as an alternative approach.

Active learning methods operate in the realm of scarce outcome-labeled data, where one has the option to interactively query an *oracle* (the user/agent in this case) for labels, and the goal is to learn the relationship between labels and relevant features using as few queries as possible (Settles 2009). These methods can help improve the efficiency of preference elicitation as well. For preference elicitation, active learning techniques can suggest new pairwise comparisons that would provide the *maximal information* about the agent's utility function given the information gathered so far (Dragone, Teso, and Passerini 2018). Using this form of structured determination of the next pairwise comparison (based on the agent's previous responses), the agent's preferences can be inferred faster than the setting where they are presented with random comparisons at each time step. For this reason, multiple recent works consider active-learning-based preference elicitation. Yang et al. (2021) use interactive elicitation to create recommendation systems. Srivastava, Heidari, and Krause (2019) develop active-learning-based surveys to elicit fairness preferences. Johnston et al. (2023) use active learning to learn preferences regarding healthcare resource allocation. These recent use cases of active learning provide evidence of its ability to efficiently elicit people's preferences. **However, the effectiveness of active learning relies on certain assumptions that may not hold in the case of moral preferences**.

Moral preferences capture a person's normative views over available actions in moral dilemmas—that is, what is the *right* thing to do when the chosen action could lead to significant harm to others, but not (or not only) to the participant themself? A popular example is the *trolley problem*, where the participant is asked which human lives should be prioritized, passengers or pedestrians (Foot 1967). Similarly, in the kidney allocation example described earlier, when asked to decide which of two patients on the kidney transplant list should get the kidney, a participant's decisions are based on patient features that they consider *morally relevant*. In these settings, when an agent expresses a preference for one patient over another, their judgment can be characterized by the underlying *utility function* they use to assign *relevance scores* to the available actions, choosing the action with the highest assigned score. Note that, despite the use of utility functions, this standard setup does not presuppose any utilitarian moral theory because it can model agents who base their decisions on non-utilitarian factors, such as past misbehaviors by patients. Modeling the participant's preferences in moral decision-making settings (e.g., by learning their underlying utility function) allows for predicting their moral judgments when presented with new dilemmas in the same setting. Therefore, these models can be useful in the development of ethical AI tools (Feffer et al. 2023).[1] However, eliciting moral preferences can be challenging, and differ from the process of eliciting other kinds of preferences.

Moral preferences concern harms to others, and differ from self-interested, economic, or material preferences, where the agent chooses the option with the highest subjective utility to self (Capraro and Perc 2021). Instead of being concerned only with the self, moral preferences are intended to be impartial (Vanberg 2008) and fair (Bicchieri and Chavez 2010). Computational modeling of these preferences can, therefore, help develop decision-aid tools that incorporate stakeholders' moral values, e.g., in applications like autonomous vehicles or biomedical situations. Unsurprisingly, the standards of expected elicitation accuracy in these domains are quite high, since inaccurate prediction of moral judgments can significantly harm the people using or affected by the decision-aid tool. For these reasons, greater attention to elicitation performance is required in moral decision-making settings.

Yet, a crucial problem in eliciting moral preferences is that they can be *unstable*, i.e., the participant's choices for the first few presented moral dilemmas might appear inconsistent with each other (Crockett 2016). The participant can also be indecisive and provide "noisy" judgments to moral dilemmas (e.g., there may be variability in their choices for similar scenarios), further complicating the elicitation process (Rehren and Sinnott-Armstrong 2022). Research on moral psychology also lacks consensus on the structure of cognitive processes that incorporate moral preferences within our judgments (Ugazio et al. 2022). Limited understanding of moral decision-making structures makes it difficult to model them computationally. All these properties

[1]A note on terminology: what we call *moral preferences* can also be described as judgments/orderings over available actions in moral dilemmas. This characterization is different than decision theory literature, which defines preference orderings over outcomes rather than actions (Arrow et al. 1996). Yet, we use the term preferences to be consistent with CS preference elicitation literature on modeling decision processes in pairwise comparison settings.

taken together make moral preference elicitation a complex task and call into question the validity of active learning as a reliable elicitation methodology.

The use of active learning for preference elicitation often presupposes that the context in question does not suffer from the above issues. Preference stability, limited variability in responses, and availability of a hypothesis class that captures the underlying utility u are common assumptions (Dragone, Teso, and Passerini 2018). An obvious question that then arises is *whether active learning still leads to efficient moral preference elicitation when these assumptions are violated.* Research from moral psychology suggests that these assumptions may specifically not hold for moral preferences. Hence, the efficacy of active learning for moral preference elicitation needs further examination.

Our Contributions. In this paper, we investigate whether active learning can be effective for moral preference elicitation, based on simulations designed to replicate the above challenges with moral preferences. Our simulations test two popular active learning paradigms, version-space-based active learning and Bayesian active learning (Section 2). Inspired by recent human subject research on properties of moral decision-making (Section 3), we consider the following challenges: (a) **preference instability**, (b) **model misspecification**, and (c) **noisy responses**. In all settings, we compare active-learning-based approaches against a standard approach that presents agents with random pairwise comparisons. We observe the following:

• **Preference instability:** Our simulations here evaluate elicitation performance when the agent's moral preference model stabilizes only after responding to a certain number of initial comparisons (Section 4.1). Specific scenarios we consider include: (1) the agent, after a few comparisons, simplifies their moral preference to reduce the decision-making effort, (2) the agent makes their preference more complex to incorporate additional information, and (3) the agent changes their preference entirely to reflect significant updates to their moral values. We observe that, in all three cases, when the number of features is small, the Bayesian active learning approach recovers well from instability and achieves higher accuracy than the random query baseline within a small number of comparisons after a preference change. However, in cases of drastic preference changes and a large number of features, both active learning approaches have similar or worse performance than the random query baseline due to their dependence on previous comparisons. *The key takeaway here is that the accuracy and efficiency of active learning depend on the expected scale of preference instability (as captured by the kind of preference change) and the complexity of the decision-making context (as captured by the number of features).*

• **Model misspecification:** Our model misspecification simulation evaluates preference elicitation performance when the agent's moral decision-making model and the model class used by the elicitation framework are different (Section 4.2). For instance, suppose the agent uses a shallow decision tree to encode preferences, but the preference elicitation framework uses the class of linear models. Here, active learning *at best* converges to the best hypothesis in the linear class but has a relatively high predictive error–as observed in our simulations. Along with agents that use tree-based models, we simulate other scenarios of model misspecification, such as scenarios where the agent uses feature interactions but the elicitation model doesn't, and scenarios where the agent and the elicitation model use different feature sets. When the extent of model misspecification is large, we observe that active learning approaches and random query baseline have similar performance. *The key takeaway here is that appropriate modeling of the agent's moral decision-making process is necessary for active learning to improve the elicitation efficiency of the framework.*

• **Noisy responses:** We also consider the setting where the agent's responses are stochastic and simulate two kinds of stochasticity: (a) *response noise*: when stochasticity in the agent's response to a pairwise comparison depends on the difference between utility assigned to each item in the pair (i.e., higher variability when utilities are close) and (b) *preference noise*: when the agent's preference model is sampled from a certain distribution. For response noise, we observe that the Bayesian active learning approach is still more efficient than the random query baseline despite noise. For preference noise, active learning is more efficient than random query baseline only when noise is *small* (e.g., when noise magnitude is small relative to the range of model parameter values). *The key takeaway here is that one needs to consider the source and impact of variability in agent responses to assess the effectiveness of active-learning-based elicitation.*

Overall, our simulations shed light on the performance of active learning for simulated moral preference elicitation tasks. We find that active learning can improve elicitation efficiency in certain settings (e.g., small-scale noise) but also reduce elicitation efficiency in other settings (e.g., large-scale preference instability). Based on these results, we emphasize the need to understand the nuances associated with the moral decision-making context in question before deploying active learning-based elicitation frameworks. Additionally, our findings can inform future human-subject studies aimed at understanding the extent to which these assumptions are violated in common moral preference elicitation tasks.[2]

Related Work. Preference elicitation methods are employed in multiple domains to create user-centered services, e.g. to create recommendation systems (Priyogi 2019), to understand consumer behaviour (Ben-Akiva, McFadden, and Train 2019), and for patient-centered decision-making in healthcare (Weernink et al. 2014). Research on preference elicitation similarly spans multiple disciplines, including computer science (Chen and Pu 2004), economics (Beshears et al. 2008), and psychology (Slovic 2020). Machine-aided elicitation has further improved learning efficiency by helping process available agent data and/or the choices they make in real and hypothetical scenarios (Soekhai et al. 2019). As mentioned earlier, similar efforts have been made in moral domains, with several applications employing elic-

[2]The technical appendix sections for this paper are available in the extended version on Arxiv.

itation frameworks to model moral preferences (Awad et al. 2018; Srivastava, Heidari, and Krause 2019; Loreggia et al. 2019; Balakrishnan et al. 2019; Sinnott-Armstrong, Skorburg et al. 2021; Johnston et al. 2023). For a general survey of moral preference elicitation methods, we recommend Feffer et al. (2023). In our work, we focus on methods that query an agent to choose between two given cases and use their responses to learn their preferences (Ben-Akiva, McFadden, and Train 2019). While pairwise comparisons are a popular elicitation technique, there are alternative approaches as well, e.g., asking agents to report their preference strength (Toubia et al. 2003), rank choices (Ali and Ronaldson 2012), participate in bidding processes (Conen and Sandholm 2001), or describe the motivations for their choices (Liscio et al. 2023, 2024).

Active learning can be used to either learn the agent's utility model or to successively present them with better recommendations (Houlsby et al. 2011; Dragone, Teso, and Passerini 2018). For the former setting learning the utility model, Huang and Luo (2016) propose active learning methods to learn marketplace consumer preferences and Srivastava, Heidari, and Krause (2019) elicit fairness preferences using active-learning-based surveys. For the latter setting of generating personalized recommendations, Elahi, Ricci, and Rubens (2014) and Yang et al. (2021) discuss active learning strategies to streamline data collection for recommendation systems. Johnston et al. (2023) use uncertainty-based active learning methods proposed by Vayanos et al. (2020) to model healthcare resource allocation preferences of survey participants. Our work focuses on learning the utility model since the eventual goal is to use the learned utility and preferences for downstream applications.

Most preference elicitation studies focus on preferences involving self-benefits, e.g., to create recommendation systems or better-personalized services. As mentioned earlier, moral preferences go beyond self-interest and explain people's normative impartial judgments. For instance, Bicchieri and Chavez (2010) show the insufficiency of monetary preferences in explaining people's fairness perceptions. Capraro and Rand (2018) discuss how social preference models can be incompatible with people's choices of equitable actions. Other experimental analyses from psychology (see Capraro and Perc (2021) for a review) provide further evidence of contrasts between moral and material preferences.

Beyond our work, certain recent papers examine the limitations of active learning in different contexts. Margatina and Aletras (2023) and Kottke et al. (2019) discuss the dependence of active learning's performance on common (but potentially unrealistic) assumptions, e.g. representative training data and equal labeling costs across cases. Active learning can also fail to outperform random query baselines when faced with distribution shifts (Snijders, Kiela, and Margatina 2023) or outliers (Karamcheti et al. 2021). Data collected using active learning is implicitly tied to the learning model and can lead to generalization issues (Lowell, Lipton, and Wallace 2019). Our work adds to this line of research, specifically questioning the applicability of active learning to moral preference elicitation.

Algorithm 1: Online preference elicitation

Input: Functions sample$(\cdot)$, $R(\cdot,\cdot)$, and fit$(\cdot,\cdot)$, N, class $\mathcal{H}$

1: $S \leftarrow \emptyset$
2: **for** $t \in \{1, \dots, N\}$ **do**
3: $\quad \mathbf{x}_t, \mathbf{x}'_t \leftarrow$ sample(S) {sample new comparison}
4: $\quad r_t \leftarrow R(\mathbf{x}_t, \mathbf{x}'_t)$ {Get agent's response}
5: $\quad S \leftarrow S \cup \{(\mathbf{x}_t, \mathbf{x}'_t, r_t)\}$
6: $\quad h_t \leftarrow$ fit$(S, \mathcal{H})$ {Learn hypothesis using dataset S}
7: **return** h_N

2 Algorithms for Preference Elicitation

The basic structure of the elicitation procedure is described in Algorithm 1. At time-step t, the agent is presented with a sampled comparison $(\mathbf{x}_t, \mathbf{x}'_t)$ and their response is recorded. Then, the algorithm finds the hypothesis h_t from class $\mathcal{H}$ which best fits the labeled comparisons recorded till time t.

The sampling step of Algorithm 1 (Step 3) can be executed either by randomly sampling a pair of input cases or by using *active learning*, whereby the chosen pair depends on the comparisons presented so far and the hypothesis class $\mathcal{H}$. We will use **RANDOM-PE** to refer to the instance of Algorithm 1 that uses random sampling. When using active learning to sample comparisons, multiple methods from prior works can be employed and we outline two popular approaches below. Longer mathematical descriptions and use cases of these methods are provided in Appendix A.

Version-Space-based Active Learning. Given a kernel-SVM decision boundary learned using available labeled data, the informativeness of any new query can be approximated using the distance of the query from the decision boundary and this heuristic can be used to generate an informative next query (Tong and Koller 2001). To implement this approach, we learn an SVM hypothesis f that best fits $((\mathbf{x}_i, \mathbf{x}'_i))_{i=1}^t$ to labels $(r_i)_{i=1}^t$ and find a comparison that is closest to f's decision boundary. We will call this approach **ACTIVE-VS-PE**.

Bayesian Active Learning. The Bayesian Active Learning with Disagreement (BALD) algorithm represents preferences using a Gaussian process with a specified kernel and chooses the next query to be the one that maximizes the mutual information between model predictions and model posterior (Houlsby et al. 2011). Implementation of this approach for Algorithm 1 requires learning a representation of the posterior corresponding to the labeled dataset $((\mathbf{x}_i, \mathbf{x}'_i, r_i))_{i=1}^t$ and then finding the pairwise comparison with high mutual information. We will call this approach **ACTIVE-BAYES-PE**. Note that the use of learned posterior is limited to the sampling step and can be independent of the learning step.

The final step of Algorithm 1 (Step 6) uses a pre-specified function fit$(\cdot,\cdot)$ to learn a hypothesis h_t from $\mathcal{H}$ that "best" simulates responses $(r_i)_{i=1}^t$ using comparisons $((\mathbf{x}_i, \mathbf{x}'_i))_{i=1}^t$. For instance, if $\mathcal{H}$ is the class of linear functions, then fit$(\cdot,\cdot)$ can implement an SVM, logistic regression, or any other linear classification training procedure (with appropriate regularization). Alternately, to rank cases based on the agent's responses (with $\mathcal{H}$ denoting the set of all rankings), the popular Bradley-Terry approach can be implemented within fit$(\cdot, \cdot)$ (Bradley 1984). The choice of $\mathcal{H}$ here depends on prior beliefs about the agent's preference model. However, a mismatch between the agent's preference model and $\mathcal{H}$ can impact the effectiveness of the framework, as we show in Section 4.2.

3 Challenges to Modeling Moral Preferences

Inspired by prior research from moral psychology, we highlight three obstacles to computationally modeling an agent's moral preferences. These obstacles are (a) change in preference after making a certain number of decisions, (b) the agent's model not being included in $\mathcal{H}$, and (c) noise in the agent's responses. We describe these challenges here, specifically focusing on prior empirical evidence for them from human subject research in the pairwise comparison setting.

Preference instability. Empirical studies in psychology provide extensive evidence that agent's preferences in unfamiliar contexts are developed as they make decisions in those contexts (Hoeffler and Ariely 1999; Ariely and Zakay 2001; Warren, McGraw, and Van Boven 2011; Dhar, Nowlis, and Sherman 1999). In these settings, the first few choices made by an agent can be unstable (i.e., their preferences can change after making some decisions) and may not reflect their eventual preferences for future decisions. Moral preferences can have similar instability and can be shaped by an agent's ongoing experience with the decision-making context (Crockett 2016; Rehren and Sinnott-Armstrong 2022; Helzer et al. 2017; Curry, Chesters, and Van Lissa 2019).

In the context of pairwise comparisons, data from Boestler et al. (2024) provides evidence of this phenomenon in the kidney allocation setting. In their study, participants are asked to participate in 10 sessions (one per day) and presented with 60 pairwise comparisons in each session. Session-specific analysis shows that, for many participants, there is significant variation in their weight distribution over the patient features across different sessions. In other words, for many participants, their underlying utility functions change from session to session. This kind of preference change can significantly impair the ability to computationally model moral preferences.

Note that we consider the instability of moral preferences over available actions and not preferences over moral values (e.g., one's value preference could be to prioritize equality in resource allocation over efficiency). Moral values do inform moral judgments and the preferences an agent has over available actions. But prior work has argued that while values are generally stable, agents can still be unstable in *applying* those values to make moral judgments – this is referred to as the "value-action gap" (Gould et al. 2023). In our setting, since we only observe the agent's moral judgments, we mainly focus on the challenge posed by the observed instability of preferences expressed through these judgments.

Model misspecification. Another challenge in the computational modeling of moral preferences is model misspecification, i.e., making incorrect/misrepresentative assumptions regarding the structure of the agent's decision-making process. A popular modeling assumption is the *additive independence* model, where we assume the utility the agent assigns to any input can be represented as a sum of the utilities assigned to individual input features (Chen and Pu 2004) (e.g., linear utility satisfies this assumption). Another common modeling assumption is the *complete information assumption*, i.e., all the information *explicitly used* by the agent to make their decision is available to the elicitation framework. Assumptions of this kind are common in active learning-based elicitation as they reduce the complexity of query generation (Yang et al. 2021; Johnston et al. 2023). They also affect the choice of $\mathcal{H}$ in Algorithm 1, e.g., if we assume additive independence and complete information, then setting $\mathcal{H}$ to be the linear class can help learn explainable representations of the agent's preferences.

However, in many situations, these assumptions do not reflect the agent's decision-making process (Pine et al. 2009; Gonzalez Sepulveda, Johnson, and Marshall 2021). Cognitive processes underlying moral decision-making are not clearly understood (Ugazio et al. 2022) and can be more complex than a linear combination of available features (Hofmann, Hoelzl, and Kirchler 2008). Both empirical and theoretical analyses of moral judgments highlight this complexity. Cohen and Ahn (2016) fit multiple kinds of linear and nonlinear models over people's responses in moral dilemmas and find that models from the exponential function family often provide the best fit. Kagan (1988) theoretically questions both the additive and independence assumption (in an article appropriately titled "The Additive Fallacy"), explaining through multiple *contrastive* examples that (a) moral status of an act cannot always be determined by the sum of weights of individual features, and (b) weight assigned to each feature can depend on the weight assigned to other features (i.e., feature interactions). As such, nonlinearity and dependence across various features can be expected in moral decision-making processes.

Noisy responses. Stochasticity in agent's choices, (specifically, changes in their responses to similar scenarios at different times) has been noted in various domains (Marley and Regenwetter 2016; Becker, DeGroot, and Marschak 1963). The same is true for moral decision-making domains, where response variability can be a result of ongoing deliberation, increased decision "difficulty", and/or increased complexity of the decision context (Sivill 2019). Boestler et al. (2024) provide concrete evidence of this phenomenon. In their kidney allocation study, participants take part in multiple sessions and six pairwise comparisons are repeated in each session. Participants' responses to the repeated comparisons provide insight into response variability, quantified by the fraction of times a participant's choice to a repeated scenario differed from their *majority choice* for this scenario. Boestler et al. (2024) observe significant response variability for certain repeated comparisons (in the range of 10-18%). Additionally, the results of Boestler et al. (2024) suggest that response variability is larger when the pairwise comparison is perceived as being more "difficult" by the participant, implying amplified stochasticity for difficult moral dilemmas.

All of these properties pose significant obstacles to the computational modeling of moral preferences. As we see in the following sections, the impact of these challenges can potentially be amplified by the use of active learning.

4 Testing the Efficacy of Active Learning

With the above challenges in mind, we next compare the performance of active learning for preference elicitation against the random baseline over simulations of these challenges.

Simulation setup. We primarily simulate agents that use linear utility functions, i.e., $u(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$, for any $\mathbf{x} \in \mathbf{X}$, given weights $\mathbf{w} \in \mathbb{R}^d$. The assumption of linear utility is quite prevalent in the preference elicitation literature (e.g., Noothigattu et al. (2018), McElfresh et al. (2021), Johnston et al. (2023)). In Section 4.2, we will also question this assumption and simulate agents that use tree-based models and linear models with feature interactions. To simulate an agent with linear utility, we sample weights $\mathbf{w}$ from the uniform distribution $\text{Unif}([-1, 1]^d)$. We run Algorithm 1 for each simulated agent, presenting them with N pairwise comparisons (ranging from 5 to 50). $\mathcal{H}$ is set to be the class of linear SVM classifiers over feature differences (with fit$(\cdot)$ performing SVM training). Hence, each h_t will contain the learned SVM weights, say $\hat{\mathbf{w}}_{h_t}$. We evaluate performance using two metrics: (i) accuracy - for a held-out collection of 1000 comparisons, measure the fraction of comparisons for which the response using weights $\hat{\mathbf{w}}_{h_N}$ matches that of the agent - and (ii) normalized distance - measure the L_2-distance between $\hat{\mathbf{w}}_{h_N}$ and $\mathbf{w}$ after normalization. For each setup, we report the mean and standard deviation of these metrics across 50 simulated agents. In the main body of the paper, we will primarily discuss the accuracy metric. Results with respect to distance are similar but deferred to Appendix C. The number of features d is varied from $\{3, \ldots, 15\}$ and each feature has range $\{1, \ldots, 10\}$ (unless specified otherwise). ACTIVE-VS-PE and ACTIVE-BAYES-PE will use a linear kernel function κ. Other implementation details are presented in Appendix B.

4.1 Preference Instability

The first challenge we discuss in Section 3 is preference instability, i.e., the agent's underlying preferences can change after making some decisions. Since the next query suggested by active learning depends on the agent's responses to comparisons presented so far, we simulate scenarios where an agent's preferences undergo changes to assess the impact of preference instability on active learning algorithms.

We assume that the agent's utility function is linear. Suppose that the agent changes their preferences once, at timestep $t_{\text{change}} \in [N]$. Let $w^{\text{pre}} \in R^d$ denote the agent's weight vector for all timesteps $t < t_{\text{change}}$ and $w^{\text{post}} \in R^d$ denote the agent's weight vector for all timesteps $t \geq t_{\text{change}}$. We simulate the following kinds of preference changes.

- **Downscale-ordered.** Agent changes their preference utility function to only use the feature to which they assigned the highest weight previously. For this agent, we sample pre-change preference $w^{\text{pre}} \sim \text{Unif}([-1, 1]^d)$ and set $I = \arg\max_i |w_i^{\text{pre}}|$. Then, for post-change preference, $w_I^{\text{post}} = w_I^{\text{pre}}$ and $w_i^{\text{post}} = 0$ for all $i \in [d] \setminus \{I\}$.
- **Downscale-random.** Agent changes their utility function to again use only one feature, but the feature is randomly selected. For this agent, we sample pre-change preference $w^{\text{pre}} \sim \text{Unif}([-1, 1]^d)$ and set I is chosen randomly from set $[d]$. Then, for post-change preference, $w_I^{\text{post}} = w_I^{\text{pre}}$ and $w_i^{\text{post}} = 0$ for all $i \in [d] \setminus \{I\}$.
- **Upscale-ordered**: Agent changes preference utility function from using just one feature to all features, with features in w^{post} having lower relative weight than the non-zero weight in w^{pre}. For this agent, sample $w^{\text{post}} \sim \text{Unif}([-1, 1]^d)$ and $I = \arg\max_i |w_i^{\text{post}}|$. Then, $w_I^{\text{pre}} = w_I^{\text{post}}$ and $w_i^{\text{pre}} = 0$ for all $i \in [d] \setminus \{I\}$.
- **Random-switch.** Agent changes to a random new preference after t_{change}. Here, we sample both weights vectors $w^{\text{pre}}, w^{\text{post}} \sim \text{Unif}[-1, 1]^d$, independently.

Downscale-ordered and **Downscale-random** model the settings where an agent changes their preference to reduce decision-making effort (Shah and Oppenheimer 2008). In certain cases, the agent can choose only to use the feature that was most important to them pre-t_{change}, which is modeled by **Downscale-ordered**. **Upscale-ordered** is a symmetric scenario where the agent instead incorporates additional features in their preference. Finally, **Random-switch** models agents who make more drastic changes to their preference, e.g. following an entirely different set of moral norms. Appendix C.1 models multiple other scenarios as well, e.g., agent downscaling/upscaling to or from random features (instead of highest weighted feature) and downscaling/upscaling to or from more than one feature. For all scenarios, we compare the accuracy achieved by ACTIVE-VS-PE and ACTIVE-BAYES-PE vs. RANDOM-PE, varying the number of features, number of comparisons, and t_{change}.

Results. The results of our simulations are presented in Figure 2. As expected, ACTIVE-VS-PE and ACTIVE-BAYES-PE always achieve higher accuracy than RANDOM-PE prior to t_{change}. Post-t_{change} performance shows how well each algorithm recovers from preference change.

Let us first look at the **Downscale-ordered** setting (plots on the top-left side of Figure 2). In this case, when the preference change occurs early (i.e., $t_{\text{change}} = 10$), ACTIVE-BAYES-PE recovers quite fast from the preference change: the accuracy of ACTIVE-BAYES-PE becomes higher than that of RANDOM-PE within 10 timesteps (on average) post-t_{change} when $d = 5$. In comparison, ACTIVE-VS-PE takes longer to recover and exceeds RANDOM-PE in accuracy. For larger t_{change}, both active learning approaches seem to recover slower and incompletely. When $d = 10$ and t_{change} is 20 or 30, we further observe that ACTIVE-BAYES-PE and ACTIVE-VS-PE have similar or even lower accuracy than RANDOM-PE for all timesteps post-t_{change}. This also implies reduced efficiency of active learning in settings with high feature complexity; for any desired level of accuracy, active learning approaches take a similar or larger number of comparisons than the random query baseline to achieve that accuracy level. In the case of **Downscale-random** setting, the performance of active learning algorithms, relative to RANDOM-PE, follows similar patterns – when $d = 5$

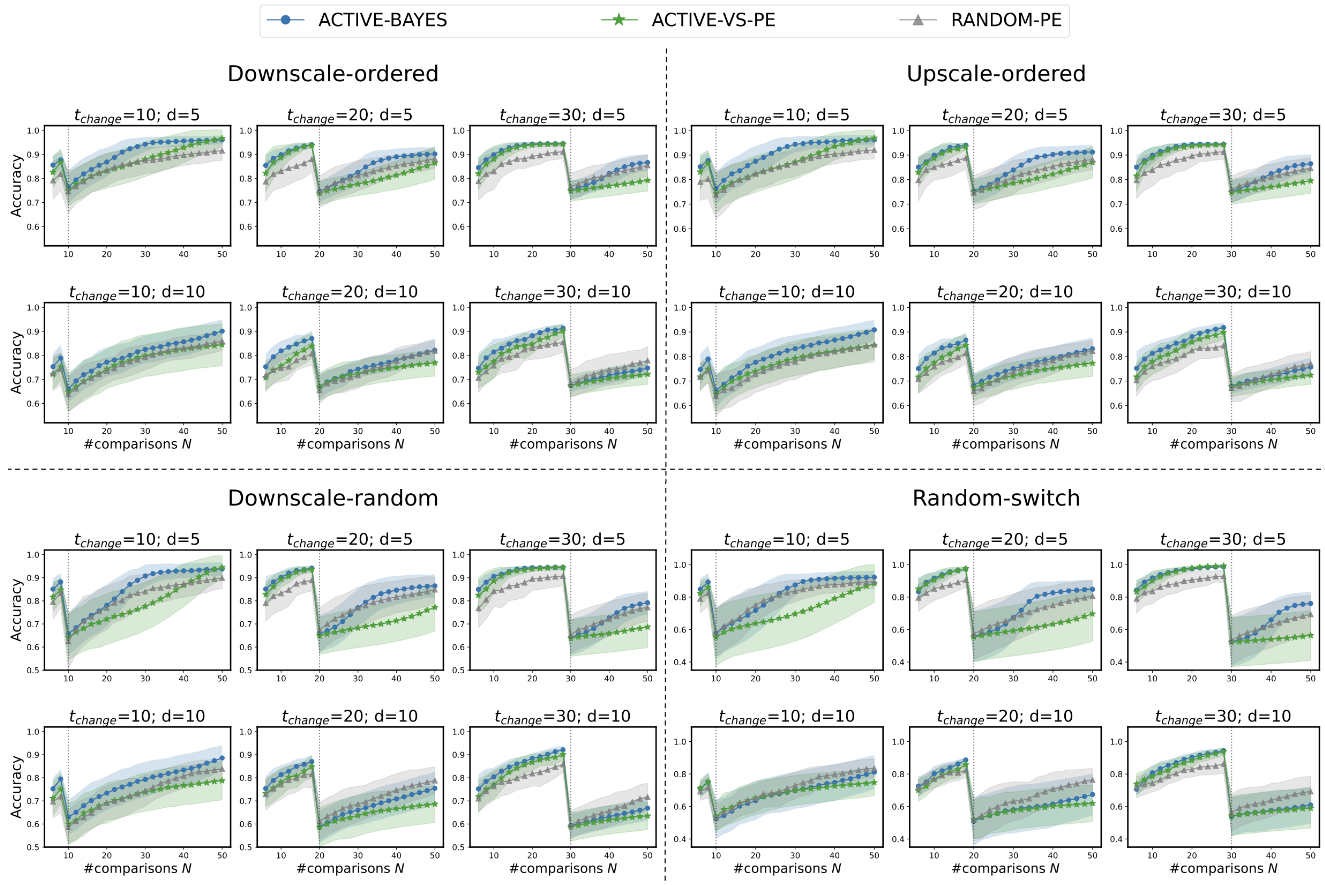

Figure 2: Performance for preference-change scenarios from Section 4.1. ACTIVE-BAYES-PE often performs better than RANDOM-PE post-t_{change} when d=5. However, in many cases (e.g., d=10, t_{change}=$20, 30$), both active learning algorithms have similar or worse performance than RANDOM-PE.

and t_{change} is small, ACTIVE-BAYES-PE recovers well compared to other algorithms, but this recovery is much slower for $d = 10$. The drop in accuracy around timestep t_{change} is also larger in magnitude for **Downscale-random** compared to **Downscale-ordered**; this is expected since there is relatively more consistency between pre-change and post-change preferences in the **Downscale-ordered** setting.

Similar trends are observed for the **Upscale-Ordered** and **Random-switch** plots in Figure 2. Active learning approaches have the worst recovery in the **Random-switch** setting where, due to the drastic change in the agent's preferences, both ACTIVE-BAYES-PE and ACTIVE-VS-PE have similar or worse performance than the RANDOM-PE post t_{change} when $d = 10$. On the positive side, when $d = 5$, ACTIVE-BAYES-PE does achieve higher accuracy than RANDOM-PE within 20 timesteps post-t_{change} on average.

Overall, Bayesian active learning approaches can efficiently elicit preferences while handling preference changes when the number of features d is small. However, these approaches fail to provide similarly improved performance as feature complexity and preference change timestep increases. These results highlight the importance of knowing the nature and scale of preference instability before deploying active learning. While active learning will eventually recover after a larger number of timesteps beyond 50, we see that in the timesteps following t_{change}, it can perform even worse than the random baseline due to its dependence on the agent's previous responses. Considering that active learning is usually employed when one has to be economical with the number of presented comparisons (due to time and/or cost constraints), not being able to rely on a certain number of initial responses can significantly affect the accuracy of the learned preferences and fail to improve, or even harm, the efficiency of the framework.

4.2 Model Misspecification

The second challenge we discussed in Section 3 is model misspecification, specifically questioning the *additive independence* and *complete information* assumptions for the agent's moral decision-making process. In this section, we evaluate active learning when additive independence and complete information assumptions are not satisfied. Setting $\mathcal{H}$ to be the linear class, we simulate the following scenarios.

- **Agent uses tree-based utility.** We simulate agents that use shallow binary decision trees to assign utility. Tree-based models reflect decisions made using *if-then* rules;

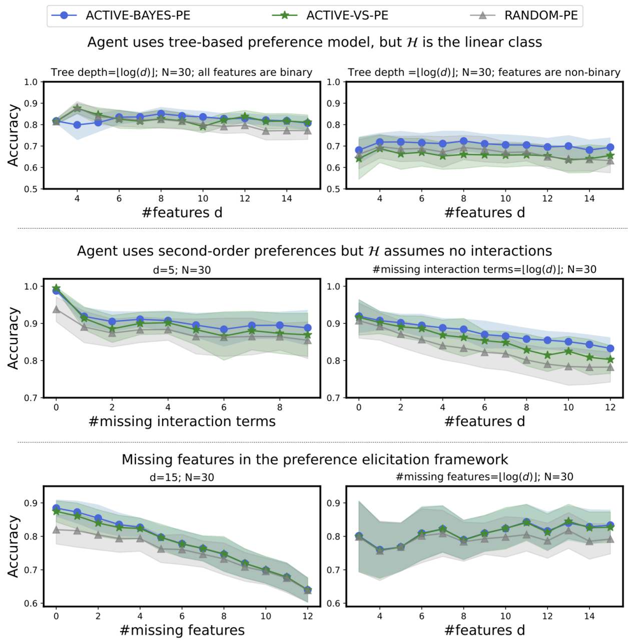

Figure 3: Performance for model misspecfication scenarios from Section 4.2. Active learning is more effective when the extent of model misspecification is small in scale.

e.g., in an organ allocation setting, an agent might assign a higher utility to a patient if their age is >50 but can be indifferent to the exact age number. To maintain parity between *capacity* of a tree model and models in $\mathcal{H}$, we simulate agents with tree models of depth $\lfloor \log d \rfloor$, where d is the number of features. We simulate this scenario with binary and non-binary features.

- **Agent uses second-order interaction terms.** Even with a linear utility model, the agent's utility function could use interactions between different features. Interaction terms account for scenarios where the importance an agent might assign to any feature is correlated with the value of another feature. For example, in the organ allocation setting, an agent might assign a higher weight to a patient's number of dependents if the patient is young, implying an interaction between the age and number of dependents variables. We simulate this scenario by measuring performance across a varying number of features d and a varying number of second-order interactions.
- **Missing features.** Finally, we consider the scenario where the agent uses information unavailable to the elicitation framework. We simulate this scenario by allowing the agent to use a larger feature set than that available for elicitation. Our simulations assess performance across a varying number of total and missing features.

Results. The results for these simulations are presented in Figure 3. When the agent uses tree-based preference, ACTIVE-BAYES-PE has marginally better accuracy than the RANDOM-PE after 30 comparisons when d is large. For small d, both active learning approaches tend to have similar or worse accuracy than the random baseline. The impact of model misspecification also depends on the input domain – overall accuracy is lower for non-binary features.

When the agent uses interaction terms, Figure 3 shows that accuracy decreases as the number of interaction terms increases. However, when the number of interaction terms is much smaller than d, ACTIVE-BAYES-PE and ACTIVE-VS-PE can achieve higher accuracy than RANDOM-PE after 30 comparisons. Finally, in the case of missing features, the larger the number of missing features (relative to d), the lower the accuracy, and the smaller the gap between ACTIVE-BAYES-PE, ACTIVE-VS-PE, and RANDOM-PE after 30 comparisons. Missing information reduces the capacity of the framework to capture the agent's decision-making process, leading to an accuracy drop.

For these scenarios, we see that the larger the scale of disparity between the agent's utility function and $\mathcal{H}$, the worse the performance of active learning as compared to the random query baseline. Active learning might still converge to the best hypothesis in $\mathcal{H}$ (see accuracy vs. timestep results in Appendix C.2); however, the above results show that disparity between functions in $\mathcal{H}$ and the agent's utility affects active learning's ability in generating informative queries and leads to a reduction in accuracy of learned preferences.

4.3 Noisy Responses

The final challenge we highlighted in Section 3 is stochasticity or variability in agent's responses to moral dilemmas. Two ways in which this stochasticity has been modeled in prior literature are (a) *response noise*: noise that arises and affects the agent's response after the agent has computed utility for the presented cases, and (b) *preference noise*: noise that arises due to variability in the agent's underlying utility function (Bhatia and Loomes 2017; Marley and Regenwetter 2016). Suppose the agent uses linear utility, i.e., $u(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$, for some $\mathbf{w}\in\mathbb{R}^d$. Then, the above noise models can simulated as follows.

- **Response noise model.** This model induces noise $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ after utility is computed. Assuming an additive noise model, the impact of this noise on the agent's response R can be interpreted as changing it to $R(\mathbf{x}, \mathbf{x}') = \mathbf{1}[u(\mathbf{x}) - u(\mathbf{x}') + \varepsilon > 0]$. Our simulations evaluate performance for varying $\sigma \in \mathbb{R}$.
- **Preference noise model.** This model assumes noise in the utility generation process itself. We simulate this setting as follows: Suppose that whenever presented with a pairwise comparison, the agent first samples $\mathbf{w} \sim \mathcal{N}(\mathbf{w}^\star, \sigma^2\mathbf{I}/d)$, and then uses the sampled $\mathbf{w}$ to compute utilities. Here, $\mathbf{w}^\star \in \mathbb{R}^d$ represents summary feature weights assigned by the agent and $\sigma \in \mathbb{R}$ is the noise parameter varied in our simulations.

Results. The results for this simulation are presented in Figure 4. As expected, increasing σ leads to a decrease in accuracy of all algorithms. However, in the case of response noise, ACTIVE-BAYES-PE has higher accuracy than the RANDOM-PE baseline even for high values of σ. Accuracy vs number of comparisons plot for $\sigma = 2$ further shows that ACTIVE-BAYES-PE starts achieving higher accuracy than RANDOM-PE with as few as 20 comparisons. Performance

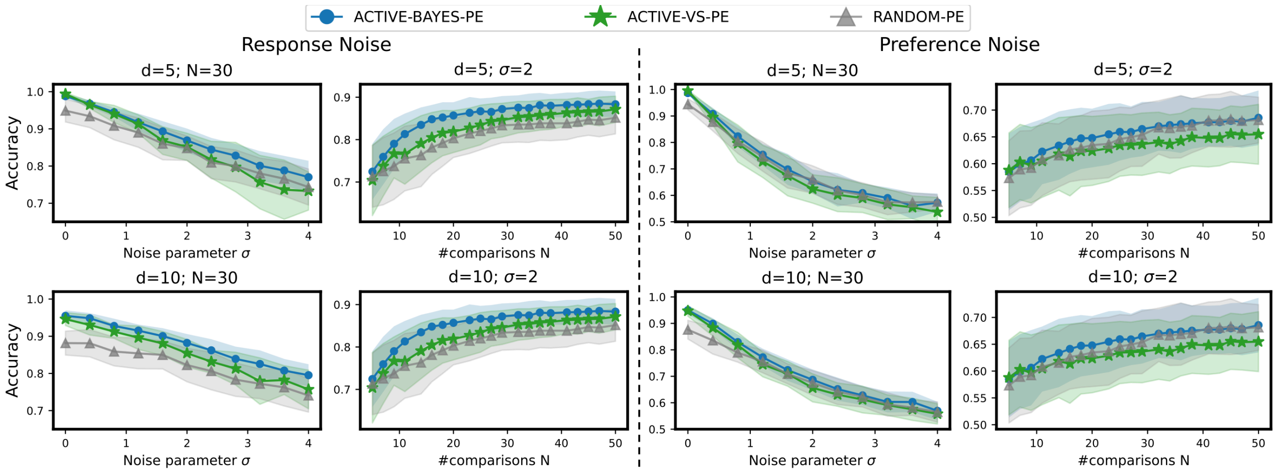

Figure 4: Performance for the noise models from Section 4.3. ACTIVE-BAYES-PE performs better than the random query baseline even with response noise. However, it fails to provide a similar improvement in most scenarios of preference noise.

of ACTIVE-VS-PE, on the other hand, is relatively better than RANDOM-PE for small σ values but becomes similar to that of RANDOM-PE for large σ. Hence, in this case, active learning (especially, ACTIVE-BAYES-PE) can be relatively more accurate at preference elicitation despite noise.

In the preference noise setting, both active learning approaches have similar performance as the RANDOM baseline for almost all non-zero σ values. Variation with respect to σ and number of comparisons shows that noise in preference weights significantly affects the ability of all algorithms to learn the underlying preferences when $\sigma > 1$. Hence, here active learning fails to provide any performance improvement in comparison to the random query generation baseline.

5 Discussion, Limitations, and Future Work

Through the presented simulations, we highlight how potential issues associated with moral preferences, such as preference instability, response variability, or modeling errors, can impact the efficacy of active-learning-based preference elicitation. In all simulated scenarios, we compare the performance of active learning-based preference elicitation against the baseline method of using random queries at each time step. Overall, there are *positive scenarios* where active learning still performs better than the random query baseline – e.g., when noise affects utility but not the underlying preferences, or in the case of small-scale preference instability in initial iterations. Then, there are *neutral scenarios* where the simulated challenge impacts the efficiency of all algorithms similarly and the performance of active learning and the random baseline are comparable – e.g., for large-scale modeling errors or when the agent's underlying preferences are noisy. In these cases, using active learning does not provide any added benefit but it also does not cause any harm to the elicitation framework. Finally, there are *negative scenarios*, where using active learning is less effective than the random baseline – e.g., when the number of features is large and the agent's preference changes after they have responded to a large number of comparisons. Here, since active learning uses the agent's previous responses to construct the next query, it takes longer to recover from preference changes.

Different real-world challenges have different effects on active learning for preference elicitation. Deploying these frameworks without prior understanding of the agent's decision-making for the given context can lead to inaccurate representations of their preferences. While using a small number of queries will almost always provide only an approximate representation of the underlying preferences, our simulations call attention to the sources of inaccuracy that were unappreciated in previous works and could lead to incorrect interpretations of results if not considered in practice.

In the paragraphs below, we highlight other characteristics of our assessment as well as future work on this topic.

Algorithmic solutions. One response to the challenges we simulate is that many of them can be addressed algorithmically if they are known in advance. If an agent's preferences are known to be unstable for initial comparisons, then one can, say, modify the elicitation approach to disregard a certain number of initial comparisons or assign sample weights to each case that are inversely proportional to the duration since the case was observed by the agent. This way, active learning can construct queries that are primarily based on the most recent agent responses. To account for feature interactions, the models in $\mathcal{H}$ can allow interactions by default and use regularization to rule out scenarios where interactions are not used. Prior work on active learning methods that are robust to noise or distribution shifts can be potentially adapted to make elicitation more resilient to noise or modeling errors (Angluin and Laird 1988; Zhao et al. 2021). In simulations, Bayesian approaches often appear more robust to certain challenges, e.g., small-scale instability. Hence, one approach is to use ACTIVE-BAYES-PE with an expanded hypothesis class $\mathcal{H}$ (e.g., combining linear and tree classes) to counter issues of model misspecification. The main challenge here is creating an efficient query-selection algorithm over an expanded $\mathcal{H}$ while being robust to instability and noise, and can be explored as part of future work. Most of these modifications, however, require prior knowledge of the nature of the challenge associated with the agent's decision-making process. Indeed, the primary goal of our analysis is to highlight that certain assumptions made when using active learning incorrectly rule out these challenges. Knowing that these assumptions might be violated can help practitioners develop modifications that might be better suited for the given context. Also, some active learning algorithms may be generally more robust to violated assumptions than others.

Sensitivity of moral preferences. As discussed, moral preferences can be different from generic preferences for self-benefit and, hence, assuming moral preferences to have a similar structure as other preferences will hurt the accuracy of the elicitation framework. Based on prior insights from the literature on moral preferences, our work discusses specific mechanisms via which these inaccuracies can occur. With the highlighted challenges and considering the emergent nature of moral psychology research, the task of eliciting moral preferences can be tricky. Nevertheless, building elicitation methods specifically for moral preferences is a worthwhile direction for future research, given their role in creating ethical AI tools. At the same time, moral preference elicitation is just one (albeit complex) part of ethical AI development. Mechanisms to incorporate learned moral preferences within AI systems involve additional work and should be similarly subjected to technical analyses of feasibility under various real-world challenges.

On utility functions. Our framework employs utility functions to model people's preferences over actions in moral dilemmas, as is standard practice in this literature. Despite the overlap in naming conventions, it is important to clarify that modeling moral preferences using utility functions does not presuppose a reliance on utilitarian or consequentialist moral theories (as long as consequentialism isn't used generically to cover all possible theories (Portmore 2022)). The justifications people have for considering features that contribute to their utility function do not have to draw on consequentialist principles, and the features people consider may not impact future consequences directly, such as when people think patients' past criminal behavior is important for determining who should receive an available kidney. Adherence to many different moral theories (including non-consequentialist theories) can be modeled using utility functions, and our analysis aims to call attention to challenges that can arise when using active learning to obtain accurate representations of various utility functions. Nevertheless, future work is needed to assess the effectiveness and challenges of using active learning to predict moral judgments under other modeling frameworks or conditions, e.g., when using explicit moral constraints (Black 2020), harm-based utilities (Beckers, Chockler, and Halpern 2022), or modified utility-based frameworks that explicitly account for deontological values (Lazar 2017).

On non-moral preferences. Issues of instability, noise, or model misspecification can arise with non-moral preferences as well. Yet, we focus on moral preferences because the specific challenges we simulate are inspired by the literature on moral philosophy and psychology. AI applications that would rely on moral preference elicitation often involve high stakes and errors in preference elicitation can cause undue harm to users and impacted individuals (e.g., in autonomous vehicles and kidney allocation settings), requiring high levels of elicitation accuracy and reliability.

Other analyses/baselines. Future assessments of active learning can also simulate violations of multiple assumptions; e.g., the presence of both preference instability and model errors. These combinations can be reflective of more complex decision-making settings. Additionally, in applications where data from past agents is available, other baselines (beyond simple random query baseline) can be considered. For instance, one could create elicitation using a curated set of queries that were informative of the preferences of past agents. All or random subsets of this curated set can be used to elicit preferences. Evaluation of active learning against such baselines can provide insight into whether it is better than methods that use prior information.

Limitations of our analysis. Our simulations demonstrate the need for improved modeling of human moral preferences and developing active learning approaches that are more robust to the simulated challenges. Along with this direction for future work, additional analyses can be conducted to further discover other failure points of quantitative preference elicitation frameworks. Note that all of our analysis simulates agents with linear or tree-based utility functions. Human moral preferences can be more complex and analyzing active learning performance through real-world data can provide more robust results. In particular, this will require human-subject studies where participants respond to comparisons generated using active learning and random comparisons. As expected, collecting this data will be expensive and time-consuming. In that regard, our simulation provides a starting point on the kind of data that can be gathered using active learning and raises challenges that need to be accounted for when analyzing this data.

6 Conclusion

The results of our simulations highlight the challenges associated with extracting accurate representations of agents' moral preferences while using as few queries as possible. In cases of large-scale instability or noise in agent preferences or responses, active learning has similar or worse performance than the random baseline. The assumptions made by the elicitation framework regarding the agent's moral preferences also impact the effectiveness of active learning. The use of active learning for moral preference elicitation therefore requires careful evaluation of modelling assumptions and the scale of expected variability in agent preferences and responses for the relevant context. If large-scale instability, noise, and/or violation of modeling assumptions are expected, then appropriate alternatives or modifications to active learning should be considered to counter such issues.

Acknowledgements

We are grateful for financial support from OpenAI & Duke.

References

Ali, S.; and Ronaldson, S. 2012. Ordinal preference elicitation methods in health economics and health services research: using discrete choice experiments and ranking methods. *British medical bulletin*, 103(1): 21–44.

Angluin, D.; and Laird, P. 1988. Learning from noisy examples. *Machine learning*, 2: 343–370.

Ariely, D.; and Zakay, D. 2001. A timely account of the role of duration in decision making. *Acta psychologica*, 108(2): 187–207.

Arrow, K. J.; Colombatto, E.; Perlman, M.; and Schmidt, C. 1996. *The Rational Foundation of Economic Behaviour*. Macmillan Press.

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.

Balakrishnan, A.; Bouneffouf, D.; Mattei, N.; and Rossi, F. 2019. Using multi-armed bandits to learn ethical priorities for online AI systems. *IBM Journal of Research and Development*, 63(4/5): 1–1.

Becker, G. M.; DeGroot, M. H.; and Marschak, J. 1963. Stochastic models of choice behavior. *Behavioral science*, 8(1): 41–55.

Beckers, S.; Chockler, H.; and Halpern, J. 2022. A causal analysis of harm. *Advances in Neural Information Processing Systems*, 35: 2365–2376.

Ben-Akiva, M.; McFadden, D.; and Train, K. 2019. Foundations of stated preference elicitation: Consumer behavior and choice-based conjoint analysis. *Foundations and Trends in Econometrics*.

Beshears, J.; Choi, J. J.; Laibson, D.; and Madrian, B. C. 2008. How are preferences revealed? *Journal of public economics*.

Bhatia, S.; and Loomes, G. 2017. Noisy preferences in risky choice: A cautionary note. *Psychological review*, 124(5): 678.

Bicchieri, C.; and Chavez, A. 2010. Behaving as expected: Public information and fairness norms. *Journal of Behavioral Decision Making*, 23(2): 161–178.

Black, D. 2020. Absolute prohibitions under risk. *Philosopher's Imprint*, 20(20).

Boestler, K.; Keswani, V.; Schaich Borg, J.; Heidari, H.; Conitzer, V.; and Sinnott-Armstrong, W. 2024. On The Stability of Moral Preferences: A Problem with Computational Elicitation Methods. *To appear in AIES 2024*.

Bradley, R. A. 1984. 14 paired comparisons: Some basic procedures and examples. *Handbook of statistics*, 4: 299–326.

Capraro, V.; and Perc, M. 2021. Mathematical foundations of moral preferences. *Journal of the Royal Society interface*, 18(175): 20200880.

Capraro, V.; and Rand, D. G. 2018. Do the right thing: Experimental evidence that preferences for moral behavior, rather than equity or efficiency per se, drive human prosociality. *Judgment and Decision Making*, 13(1): 99–111.

Chen, L.; and Pu, P. 2004. Survey of preference elicitation methods. Technical report.

Cohen, D. J.; and Ahn, M. 2016. A subjective utilitarian theory of moral judgment. *Journal of Experimental Psychology: General*, 145(10): 1359.

Conen, W.; and Sandholm, T. 2001. Preference elicitation in combinatorial auctions. In *Proceedings of the 3rd ACM Conference on Electronic Commerce*, 256–259.

Crockett, M. J. 2016. Computational modeling of moral decisions. In *The social psychology of morality*, 71–90. Routledge.

Curry, O. S.; Chesters, M. J.; and Van Lissa, C. J. 2019. Mapping morality with a compass: Testing the theory of 'morality-as-cooperation'with a new questionnaire. *Journal of Research in Personality*, 78: 106–124.

Dhar, R.; Nowlis, S. M.; and Sherman, S. J. 1999. Comparison effects on preference construction. *Journal of consumer research*, 26(3).

Dragone, P.; Teso, S.; and Passerini, A. 2018. Constructive preference elicitation. *Frontiers in Robotics and AI*, 4: 71.

Elahi, M.; Ricci, F.; and Rubens, N. 2014. Active learning strategies for rating elicitation in collaborative filtering: a system-wide perspective. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1): 1–33.

Feffer, M.; Skirpan, M.; Lipton, Z.; and Heidari, H. 2023. From Preference Elicitation to Participatory ML: A Critical Survey & Guidelines for Future Research. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*.

Foot, P. 1967. *The problem of abortion and the doctrine of double effect*, volume 5. Oxford.

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.

Gonzalez Sepulveda, J. M.; Johnson, F. R.; and Marshall, D. A. 2021. Incomplete information and irrelevant attributes in stated-preference values for health interventions. *Health Economics*, 30(11): 2637–2648.

Gould, R. K.; Soares, T. M.; Arias-Arévalo, P.; Cantú-Fernandez, M.; Baker, D.; Eyster, H. N.; Kwon, R.; Prox, L.; Rode, J.; Suarez, A.; et al. 2023. The role of value (s) in theories of human behavior. *Current Opinion in Environmental Sustainability*, 64: 101355.

Helzer, E. G.; Fleeson, W.; Furr, R. M.; Meindl, P.; and Barranti, M. 2017. Once a utilitarian, consistently a utilitarian? Examining principledness in moral judgment via the robustness of individual differences. *Journal of personality*, 85(4): 505–517.

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.

Hofmann, E.; Hoelzl, E.; and Kirchler, E. 2008. A comparison of models describing the impact of moral decision making on investment decisions. *Journal of Business Ethics*, 82: 171–187.

Houlsby, N.; Huszár, F.; Ghahramani, Z.; and Lengyel, M. 2011. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*.

Huang, D.; and Luo, L. 2016. Consumer preference elicitation of complex products using fuzzy support vector machine active learning. *Marketing Science*, 35(3): 445–464.

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.

Kagan, S. 1988. The additive fallacy. *Ethics*, 99(1): 5–31.

Kahng, A.; Lee, M. K.; Noothigattu, R.; Procaccia, A.; and Psomas, C.-A. 2019. Statistical foundations of virtual democracy. In *International Conference on Machine Learning*, 3173–3182. PMLR.

Karamcheti, S.; Krishna, R.; Fei-Fei, L.; and Manning, C. D. 2021. Mind Your Outliers! Investigating the Negative Impact of Outliers on Active Learning for Visual Question Answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Kottke, D.; Schellinger, J.; Huseljic, D.; and Sick, B. 2019. Limitations of assessing active learning performance at runtime. *arXiv preprint arXiv:1901.10338*.

Lazar, S. 2017. Deontological decision theory and agent-centered options. *Ethics*, 127(3): 579–609.

Lee, M. K.; Kusbit, D.; Kahng, A.; Kim, J. T.; Yuan, X.; Chan, A.; See, D.; Noothigattu, R.; Lee, S.; Psomas, A.; et al. 2019. WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on human-computer interaction*, 3(CSCW): 1–35.

Liscio, E.; Lera-Leri, R.; Bistaffa, F.; Dobbe, R. I.; Jonker, C. M.; Lopez-Sanchez, M.; Rodriguez-Aguilar, J. A.; and Murukannaiah, P. K. 2023. Inferring Values via Hybrid Intelligence. In *HHAI 2023: Augmenting Human Intellect*, 373–378. IOS Press.

Liscio, E.; Siebert, L. C.; Jonker, C. M.; and Murukannaiah, P. K. 2024. Value Preferences Estimation and Disambiguation in Hybrid Participatory Systems. *arXiv preprint arXiv:2402.16751*.

Loreggia, A.; Mattei, N.; Rossi, F.; and Venable, K. B. 2019. Metric Learning for Value Alignment. In *AISafety@ IJCAI*.

Lowell, D.; Lipton, Z. C.; and Wallace, B. C. 2019. Practical Obstacles to Deploying Active Learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 21–30.

Margatina, K.; and Aletras, N. 2023. On the Limitations of Simulating Active Learning. *arXiv preprint arXiv:2305.13342*.

Marley, A.; and Regenwetter, M. 2016. Choice, preference, and utility: Probabilistic and deterministic representations. *New handbook of mathematical psychology*, 1: 374–453.

McElfresh, D.; Chan, L.; Doyle, K.; Sinnott-Armstrong, W.; Conitzer, V.; Borg, J. S.; and Dickerson, J. 2021. Indecision Modeling. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*.

Noothigattu, R.; Gaikwad, S.; Awad, E.; Dsouza, S.; Rahwan, I.; Ravikumar, P.; and Procaccia, A. 2018. A voting-based system for ethical decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Pine, A.; Seymour, B.; Roiser, J. P.; Bossaerts, P.; Friston, K. J.; Curran, H. V.; and Dolan, R. J. 2009. Encoding of marginal utility across time in the human brain. *Journal of Neuroscience*, 29(30): 9575–9581.

Portmore, D. W. 2022. Consequentializing. *Stanford Encyclopedia of Philosophy*.

Priyogi, B. 2019. Preference elicitation strategy for conversational recommender system. In *Proceedings of the twelfth ACM international conference on web search and data mining*, 824–825.

Rehren, P.; and Sinnott-Armstrong, W. 2022. How Stable are Moral Judgments? *Review of Philosophy and Psychology*.

Settles, B. 2009. Active learning literature survey.

Shah, A. K.; and Oppenheimer, D. M. 2008. Heuristics made easy: an effort-reduction framework. *Psychological bulletin*, 134(2): 207.

Sinnott-Armstrong, W.; Skorburg, J. A.; et al. 2021. How AI can AID bioethics. *Journal of Practical Ethics*, 9(1).

Sivill, T. 2019. Ethical and statistical considerations in models of moral judgments. *Frontiers in Robotics and AI*, 6: 449751.

Slovic, P. 2020. The construction of preference. In *Shaping entrepreneurship research*. Routledge.

Snijders, A.; Kiela, D.; and Margatina, K. 2023. Investigating Multi-source Active Learning for Natural Language Inference. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2179–2201.

Soekhai, V.; Whichello, C.; Levitan, B.; Veldwijk, J.; Pinto, C. A.; Donkers, B.; Huys, I.; van Overbeeke, E.; Juhaeri, J.; and de Bekker-Grob, E. W. 2019. Methods for exploring and eliciting patient preferences in the medical product lifecycle: a literature review. *Drug discovery today*, 24(7): 1324–1331.

Srivastava, M.; Heidari, H.; and Krause, A. 2019. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2459–2468.

Tong, S.; and Koller, D. 2001. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov): 45–66.

Toubia, O.; Simester, D. I.; Hauser, J. R.; and Dahan, E. 2003. Fast polyhedral adaptive conjoint estimation. *Marketing Science*, 22(3).

Ugazio, G.; Grueschow, M.; Polania, R.; Lamm, C.; Tobler, P.; and Ruff, C. 2022. Neuro-computational foundations of moral preferences. *Social Cognitive and Affective Neuroscience*, 17(3): 253–265.

Vanberg, V. J. 2008. On the economics of moral preferences. *American journal of Economics and Sociology*, 67(4): 605–628.

Vayanos, P.; Ye, Y.; McElfresh, D.; Dickerson, J.; and Rice, E. 2020. Robust active preference elicitation. *arXiv preprint 2003.01899*.

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.

Weernink, M. G.; Janus, S. I.; Van Til, J. A.; Raisch, D. W.; Van Manen, J. G.; and IJzerman, M. J. 2014. A systematic review to identify the use of preference elicitation methods in healthcare decision making. *Pharmaceutical medicine*, 28: 175–185.

Yang, H.; Sanner, S.; Wu, G.; and Zhou, J. P. 2021. Bayesian Preference Elicitation with Keyphrase-Item Coembeddings for Interactive Recommendation. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*.

Zhao, E.; Liu, A.; Anandkumar, A.; and Yue, Y. 2021. Active learning under label shift. In *International Conference on Artificial Intelligence and Statistics*. PMLR.