Adaptive Pluralistic Alignment: A pipeline for dynamic artificial democracy

Rachel Freedman [1]

Abstract

Prevailing alignment methods target a fixed set of preferences and therefore risk forcing *value lock-in* as societal norms evolve over time. We introduce *Adaptive Pluralistic Alignment* (APA), a modular pipeline for updating pluralistically aligned AI systems to track evolving values and avoid value lock-in without repeating costly pre-training or large-scale data collection. APA has three stages: (1) learning compact personalized reward models via low-rank reward basis decomposition, (2) using these models as a jury that collectively selects among candidate outputs through social-choice-theoretic voting, and (3) efficiently adapting the jury over time by fitting new annotator weights over the fixed reward bases as values shift. The resulting system is efficient, explainable, steerable, and modular. We implement a proof-of-concept instantiation using the PRISM multi-user alignment dataset and simulated historical annotators, and provide preliminary analysis showing that jury composition and the choice of voting rule can substantially affect outcomes, particularly when jury preferences are heterogeneous. We provide full code and resulting preference datasets at `github.com/RachelFreedman/APA`.

1. Introduction

Modern AI systems increasingly mediate decisions on behalf of societies of diverse stakeholders. AI models are deployed to moderate online communities (Sparrow et al., 2024), to support consensus-building on contested public questions (Bakker et al., 2022; Jarrett et al., 2025), and to inform allocation and classification decisions in public-benefit domains such as organ exchange, food donation, and social services (Freedman et al., 2020; Lee et al., 2019; Chouldechova et al., 2018). As the scope of these deployments grows, aligning systems to any single annotator or canonical value set risks over-representing dominant groups and entrenching existing inequities. A growing body of work on *pluralistic alignment* therefore asks how AI systems can represent, rather than collapse, the heterogeneity of human values (Sorensen et al., 2024; Conitzer et al., 2024; Kirk et al., 2024b). We argue that pluralistic alignment requires explicit, steerable aggregation of stakeholder preferences, rather than the implicit aggregation performed by training a single reward model on pooled feedback. When aggregation is explicit, the outputs can be audited, debated, and adjusted to ensure fair and just outcomes.

However, even pluralistically value aligned models may become unaligned as societal values evolve over time. If the values encoded in AI systems don't adapt to these changes, the AI systems will diverge from societal consensus and discourage further societal moral development, causing *value lock-in* (Weidinger et al., 2021; Finnveden et al., 2022). Imagine that the Dartmouth Summer Research Project succeeded in creating AI in 1956, aligned it to their mid-20th-century values, and utilized it as widely as we utilize LLMs today. That system's values would be regarded as repressive in the 21st century, and its continued deployment would entrench the past moral consensus as an ongoing constraint on our behavior. The obvious solution to value lock-in is to repeat the full alignment pipeline whenever societal values change. But this won't always be economically feasible, since large-scale preference collection is expensive, and frontier-model training costs are growing rapidly. This imposes a prohibitive *alignment tax*: if the cost of keeping models aligned is too high, developers will defer or skip realignment, and value lock-in will persist. A practical pluralistic alignment method must therefore be able to track evolving values without regathering data or retraining from scratch at every update.

We introduce *Adaptive Pluralistic Alignment* (APA), an end-to-end pipeline that extends reward modeling and preference aggregation techniques to handle value change over time while reusing the expensive components of the initial training run. APA has three stages (summarized in Figure 1):

[1]Department of Electrical Engineering and Computer Science, University of California, Berkeley, USA. Correspondence to: Rachel Freedman <rachel.freedman@berkeley.edu>.

Pluralistic Alignment Workshop @ ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

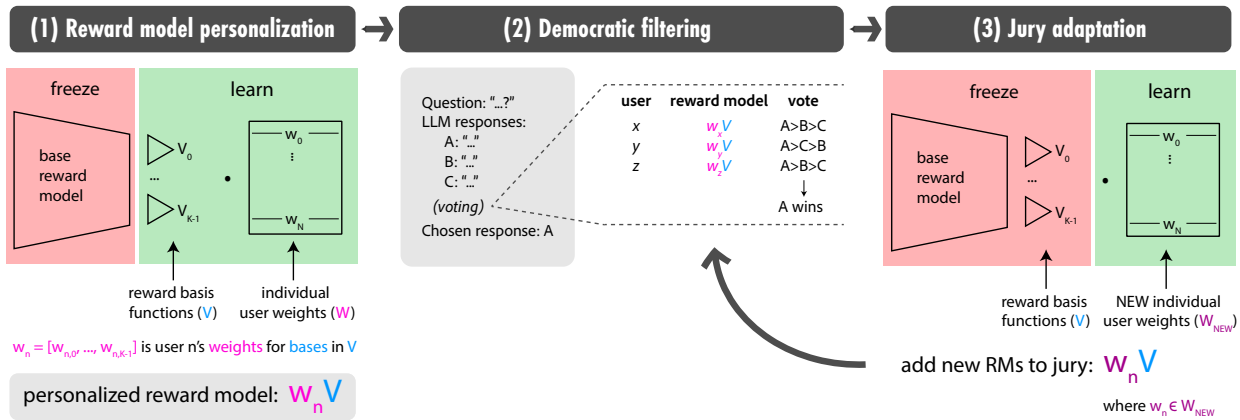

Figure 1. The full *Adaptive Pluralistic Alignment* (APA) pipeline. *(Stage 1)* **Reward model personalization**: Given multi-user preference comparison dataset $\mathcal{D}$ and embeddings from a base reward model, simultaneously learn reward basis functions V and set of individual user weights W. Each user n has weights $w_n \in W$ and personalized reward model (RM) $R_n = w_n V$. *(Stage 2)* **Democratic filtering**: At inference time, sample a distribution of responses from the LLM and construct a jury of personalized reward models. Each reward model ranks the responses, then the rankings are aggregated. The LLM gives the winning response. *(Stage 3)* **Jury adaptation**: Given a new, smaller multi-user dataset $\mathcal{D}_{NEW}$, learn weights W_{NEW} for new users. Keep V fixed. At future inferences, repeat Stage 2, now including reward models from these new users in the jury.

1. **Reward model personalization**: We apply low-rank reward modeling (Bose et al., 2025; Shenfeld et al., 2025; Barreto et al., 2026; Park et al., 2024; Zhong et al., 2024) to learn a small set of reward basis functions that span the preference variation in a given population. Individual stakeholders' preferences are described by linear combinations of these bases, making individual reward models cheap to store, fast to evaluate, and efficient to learn.

2. **Democratic filtering**: We select a subset of these individual reward models to act as a "jury" for filtering AI decisions. At inference time, we prompt the AI to generate a diverse set of decision candidates, use each jury reward model to generate a full ranking over the decisions, and then select a single winning decision using democratic voting rules from social choice theory (Conitzer et al., 2024; Ge et al., 2024; Dai & Fleisig, 2024). Using an explicit aggregation procedure makes the final decision selection steerable and auditable.

3. **Jury adaptation**: As time passes and values evolve, we periodically re-gather preferences from a new set of participants using a small, carefully chosen subset of the original questions. We hold the reward basis fixed and learn a new set of linear weights for each new participant. Because only participant weights are relearned, adaptation is orders of magnitude cheaper than retraining. We ensure that these new reward models are represented on subsequent juries when evaluating AI decisions.

Adaptive pluralistic alignment has four properties that matter for real-world deployment. The aggregation step is *steerable* and *explainable*: because votes are cast by identifiable stakeholder models under an auditable rule with known properties, the mapping from individual preferences to collective output can be inspected and adjusted, rather than buried inside a single pooled reward model. The method is inherently *robust to reward hacking*: because the final signal aggregates across a diverse jury of reward models, policies cannot exploit idiosyncratic flaws in any single model without forfeiting support from the others. The pipeline is *modular*, constructed of independent components (reward basis learning, preference aggregation, jury selection, active-query strategy) that can be swapped out to take advantage of future research improvements in each of these areas. Finally, the overall method imposes only a *minimal alignment tax* on top of existing RLHF pipelines, since the bulk of the compute is spent once on learning the bases, while per-interval updates require only a small batch of targeted queries.

Our contributions are as follows:

1. We identify *adaptive pluralistic alignment* – the problem of updating a pluralistically aligned system to track evolving societal values without repeating costly pretraining, finetuning, or large-scale reward modeling – as a distinct problem within the pluralistic alignment agenda.

2. We propose a practical, end-to-end pipeline combining personalized reward modeling, inference-time democratic filtering, and targeted jury adaption to capture societal value change.

3. We provide a proof-of-concept implementation of the pipeline, demonstrate it on a pre-existing pluralistic

alignment dataset (PRISM (Kirk et al., 2024b)), and provide preliminary analysis of how it adapts to value change over time.

This paper introduces APA as a research direction and presents a proof-of-concept implementation intended to make the framework concrete and surface design questions for follow-up work. The proof-of-concept implementation in Section 5 is illustrative rather than evaluative, and we leave systematic empirical analysis to follow-up work.

2. Related Work

Pluralistic alignment Language models reflect annotator biases and systematically underrepresent certain demographic and ideological groups (Santurkar et al., 2023). The emerging field of *pluralistic alignment* therefore seeks to align AI systems to the diverse values of the groups they serve, rather than to a single canonical preference (Sorensen et al., 2024). Sorensen et al. (2024) identify three complementary approaches: *Overton* pluralism, which aims for high coverage of the range of reasonable viewpoints, *distributional* pluralism, which seeks to represent the population-level distribution of preferences, and *steerable* pluralism, which enables users or operators to faithfully adjust the system's expressed values. In this work, we extend pluralistic alignment across time, arguing that pluralistically aligned models must also adapt to dynamic shifts in societal values. We call this problem *adaptive pluralistic alignment* (APA). APA is a fairly new and under-studied direction: the closest prior work that we are aware of is ProgressGym (Huang et al., 2024), which frames the related problem of extrapolating the forward trajectory of moral development as "progress alignment" and attempts to model the generative process of moral change. In contrast, APA simply adapts to new preference data *after value change has occurred*: a simpler problem that permits more practical solutions and is agnostic to the direction of time, allowing us to evaluate our pipeline on historical data.

Reward basis decomposition One popular set of approaches for learning personalized reward models is to decompose the space of human preferences into a small set of shared basis functions (Bose et al., 2025; Shenfeld et al., 2025; Barreto et al., 2026; Park et al., 2024; Zhong et al., 2024). These approaches learn a set of K reward basis functions from a large multi-annotator dataset, representing each individual annotator's reward model as a linear combination of these bases parameterized by a K-dimensional weight vector. Individual reward models are therefore compact, fast to evaluate, and cheap to learn for new users, since fitting a new annotator requires estimating only K weights rather than training an entire reward model from scratch. Shenfeld et al. (2025) demonstrate that as few as ten preference comparisons suffice to infer user-specific weights, and Barreto et al. (2026) provide PAC-style generalization guarantees showing how approximation error scales with the number of raters and training examples. Although the learned bases are not inherently interpretable, the low-rank structure imposes a useful inductive bias by forcing the model to discover shared axes of preference variation. We take advantage of the compactness and data efficiency of this approach to enable rapid adaptation in our pipeline. By learning reward bases once from a large initial dataset and holding them fixed, we can cheaply incorporate new annotators at future time intervals by fitting only their weight vectors, rather than repeating the expensive basis-learning step.

Preference aggregation A growing body of work applies concepts and methods from social choice theory (SCT) to the problem of aggregating diverse human feedback for AI alignment. Conitzer et al. (2024) argue that the aggregation step implicit in standard RLHF – pooling feedback from heterogeneous annotators into a single reward model – is itself a social choice problem. Siththaranjan et al. (2023) formalize this intuition by proving that standard RLHF implicitly aggregates annotator preferences according to Borda count, a voting rule that can systematically underweight minority preferences and produce counterintuitive outcomes. Subsequent work has studied the axiomatic properties of various alignment aggregation procedures (Ge et al., 2024; Gölz et al., 2025), explored policy-level aggregation (Alamdari et al., 2024), and investigated the learnability of social choice functions from data (Pardeshi et al., 2024). Our pipeline treats preference aggregation as a modular component: any social choice function can be substituted into the aggregation step without altering the reward modeling or adaptation stages. For our proof-of-concept implementation, we use the `pref_voting` library (Holliday & Pacuit, 2025), which provides implementations of a wide range of voting rules. This modularity allows our system to incorporate future advances in preference aggregation as they emerge.

3. Preliminaries

Reward learning The goal of *reward learning* is to estimate a function $\mathcal{R} : \mathcal{I} \to \mathbb{R}$ mapping alternatives from the set $\mathcal{I}$ to scalar reward values. We focus on the most common type of reward learning, *preference learning*, in which the reward function is inferred from human preference comparisons between pairs of alternatives. Because human annotators can make mistakes, they are typically modeled as noisily rational, choosing alternative i over j with probability $\Pr(i \succ j)$ rather than deterministically. In the reward learning literature, human feedback is most often modeled as Boltzmann-rational (Bradley & Terry, 1952; Jeon et al., 2020), where the probability that a teacher with

rationality parameter $\beta \in [0, \infty)$ prefers item i to j is:

$$\Pr(i \succ j; \beta, \mathcal{R}) = \frac{\exp(\beta \mathcal{R}(i))}{\exp(\beta \mathcal{R}(i)) + \exp(\beta \mathcal{R}(j))}, \quad (1)$$

where $\mathcal{R} : \mathcal{I} \to \mathbb{R}$ gives the true reward value of all items in set $\mathcal{I}$. Boltzmann-rationality does not fully model all nuances of human decision-making (Lindner & El-Assady, 2022), but it does capture the important property that humans are more likely to make mistakes when $|\mathcal{R}(i) - \mathcal{R}(j)|$ is small and therefore the comparison is harder (Barnett et al., 2023), which likely accounts for its empirical success and pervasive use in practice. It is now the dominant model for safety finetuning large language models.

Low-rank reward modeling (LoRe) A recent line of work observes that the diversity of preferences across a population of annotators can be captured in a low-dimensional subspace (Bose et al., 2025; Shenfeld et al., 2025; Barreto et al., 2026; Park et al., 2024; Zhong et al., 2024). Low-rank reward modeling (LoRe) (Bose et al., 2025) exploits this to learn compact personalized reward models. LoRe learns a set of K *basis functions* $V = [v_1, v_2, \ldots, v_K]$, implemented as reward heads sharing a common backbone, that together span the space of preference variation in a heterogeneous annotator population. Each annotator n is then represented by a weight vector $w_n \in \mathbb{R}^K$, and their personalized reward model is given by the linear combination $\mathcal{R}_n = w_n V$. Given a dataset $\mathcal{D}$ of pairwise preference comparisons labeled with annotator identities, the bases V and annotator weight matrix $W \in \mathbb{R}^{K \times N}$ are learned jointly by maximizing the likelihood of the observed comparisons under the Bradley-Terry model (Equation 1). Crucially, once the bases are fixed, fitting a new annotator requires learning only a K-dimensional weight vector, which can be done from very few preference comparisons. This compactness and data efficiency make LoRe well suited to our setting: we learn the bases once from a large initial dataset and subsequently reuse them to cheaply add new annotators to our jury at future time intervals.

Social choice functions A *social choice function* (SCF) takes as input a collection of individual preference orderings over a set of alternatives and returns a winning alternative (Dai & Fleisig, 2024; Conitzer et al., 2024). Formally, let $\mathcal{A} = \{a_1, \ldots, a_M\}$ be a set of alternative candidates (in our setting, possible AI decisions), and let $\succ_n$ denote the full preference ordering of annotator n over $\mathcal{A}$. The full *preference profile* is the set of all individual preference orderings $\sigma = (\succ_1, \ldots, \succ_N)$. A SCF $\sigma \mapsto a^*$ aggregates these and chooses a single winner $a^* \in \mathcal{A}$. Different SCFs satisfy different axiomatic properties, and impossibility results show that no single rule can satisfy all desirable properties simultaneously (Mishra, 2023; Ge et al., 2024). In practice, the choice of SCF involves trade-offs among properties such as *Condorcet consistency* (always selecting a candidate that beats every other in pairwise majority), *independence of clones* (immunity to the introduction of near-duplicate alternatives), and *resistance to strategic voting*. Which properties matter most depends on the deployment context. Because our pipeline treats the aggregation step as a modular component, any SCF can be substituted without altering the rest of the system.

4. Adaptive pluralistic alignment

The *adaptive pluralistic alignment* (APA) pipeline has three stages: (1) reward model personalization, (2) democratic filtering, and (3) jury adaptation (summarized in Figure 1). Here we formally define each stage, describe our proof-of-concept implementations, and outline directions for future improvement.

4.1. Stage 1: Reward model personalization

The first step is to learn a set of personalized reward models representing our potential jury members. We assume access to a dataset $\mathcal{D}_0$ of multi-annotator preference comparisons gathered at the initial time $t = 0$, in which each item $(i, j, n) \in \mathcal{D}_0$ represents annotator n's preference over the two outcomes i, j. In practice, we use the PRISM alignment dataset (Kirk et al., 2024a), which surveyed 1,500 demographically and geographically diverse participants for their preferences over LLM responses in 2023.

We use the LoRe method (Bose et al., 2025) to simultaneously fit both a set of K reward basis functions (V) with parameters θ that explain the diversity in annotator preferences and linear weights ($W \in \mathbb{R}^{N \times K}$) for each of the N annotators in $\mathcal{D}_0$, where the linear product WV^θ gives a vector of all N annotator reward functions. The training objective is to minimize the negative log-likelihood:

$$\min_{\theta \in \Theta, \{w_n \in \Delta^{K-1}\}_{n \in N}} \sum_{(i,j,n) \in \mathcal{D}_0} - \log\big(\Pr(i \succ j;\, \beta = 1,\, \mathcal{R} = w_n V^\theta)\big) \quad (2)$$

where $\Pr(i \succ j; \beta, \mathcal{R})$ is given by Equation 1. Following Bose et al. (2025), we constrain $w_n \in \Delta^{K-1}$ and fix $\beta = 1$. In practice, we implement $K = 8$ basis functions as shallow heads on an existing open-source reward model (`Skywork-Reward-Llama` (Liu et al., 2024)). We filter $\mathcal{D}_0$ for the 1029 annotators that each have at least 5 preferences in the dataset, then jointly learn W and θ. The N annotators with weights in W form our initial pool $\mathbb{J}$ of potential jury members. Figure 2 shows learned weights for a subset of the jurors, with weights learned from PRISM users in blue.

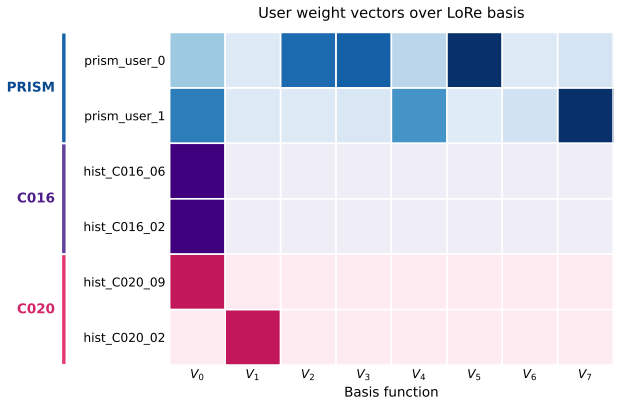

Figure 2. Subset of jury members, represented by weights over the $K = 8$ basis functions. Blue jury members are learned from real user data in Stage 1 (PRISM), while purple and pink jury members are learned from simulated historical user data in Stage 3.

4.2. Stage 2: Democratic filtering

At inference time, we use the personalized reward models to collectively select among candidate responses. This stage requires two inputs: a set $\mathcal{A}$ of candidate response alternatives and a jury $J \subseteq \mathbb{J}$ of personalized reward models drawn from the pool of potential jurors.

Ideally, the candidate set $\mathcal{A}$ should be *diverse*, covering a wide range of potentially preferred outcomes so that the jury has meaningful choices to adjudicate. Similarly, the jury should be *representative* of the stakeholder population affected by the decision. However, how best to operationalize diversity and representativeness in this setting is an active area of research with many possible desiderata. For example, we may want the distribution of alternatives to satisy *Overton pluralism*, ensuring that the full spectrum of reasonable viewpoints is represented among the candidates (Sorensen et al., 2024). We may want the jury composition to satisfy *justified representation*, ensuring that sufficiently large stakeholder groups are guaranteed a voice in the outcome (De et al., 2026). For our proof-of-concept implementation, we keep both selection procedures simple: we prompt the LLM for diverse alternatives and manually filter the output, and we select jury members randomly while ensuring equal representation from each time period. We leave the development and incorporation of more principled selection methods to future work.

The jury J votes on the candidate set $\mathcal{A}$ to determine the system output. Specifically, each juror $n \in J$ ranks the elements of $\mathcal{A}$ according to their personalized reward model $R_n = w_n V^\theta$, producing a preference ordering $\succ_n$. The collection of all jurors' orderings forms a preference profile $\sigma = (\succ_n)_{n \in J}$. We apply a social choice function to σ to determine the winning alternative a^*, which the pipeline returns as its output.

The choice of SCF can significantly impact the pipeline's behavior. There is a rich literature analyzing the axiomatic and computational trade-offs among voting rules, and future work should more precisely characterize which properties are most important in the APA setting. For our experiments, we use *Instant Runoff Voting* with parallel-universe tie-breaking (IRV-PUT) as our primary SCF, because it satisfies *independence of clones*: if a near-duplicate of an existing candidate is added to $\mathcal{A}$, the outcome does not change. This property is particularly important in our setting, since LLM text generation tends to produce clusters of similar responses that would otherwise split support among like-minded jurors. For comparison, we also report results using *Copeland's method*, a Condorcet-consistent rule that selects the candidate winning the most pairwise majority contests; *Borda count*, the aggregation rule that standard RLHF performs implicitly (Siththaranjan et al., 2023); and *plurality*, the familiar first-past-the-post rule included as a simple baseline.

4.3. Stage 3: Jury adaptation

As societal values evolve over time, the jury pool must be updated to reflect more contemporary perspectives. We accomplish this by collecting a small set of new preference data and using it to learn additional personalized reward models, which are then added to the pool of potential jurors.

Concretely, suppose that at some later time $t > 0$, sufficient time has elapsed that we expect societal values to have shifted. We subsample a set of preference comparison questions from the original dataset $\mathcal{D}_0$ and present them to a new group of participants, obtaining a smaller dataset $\mathcal{D}_t$ of preference comparisons for the new population. We then apply LoRe to learn a new set of annotator weights W_t over the *fixed* basis functions V^θ learned in Stage 1. Because the bases are held constant and only lightweight weight vectors are fitted, this adaptation step is fast and data-efficient relative to the basis-fitting procedure. The new annotator weights W_t are added to the juror pool $\mathbb{J}$, making these annotators available for selection onto future juries. Subsequent decisions are made by repeating Stage 2 with juries that may now include members from this updated pool. This process can be repeated at each successive time interval $t = 1, 2, \ldots$ as values continue to evolve, incrementally expanding the juror pool without retraining the reward bases.

The choice of which questions to subsample from $\mathcal{D}_0$ is consequential because the selected questions must capture the dimensions along which values have changed. For our proof-of-concept implementation, we use a combination of LLM-based filtering and manual inspection to select questions on topics where we expect values to vary across centuries (e.g. gender roles, religion). However, future work should draw on the active learning literature to identify the

most informative questions.

5. Proof-of-concept demonstration

We illustrate the APA pipeline with a worked example, using historical-period simulations to stand in for the future value shifts that APA is ultimately designed to track. Our goal is to make the pipeline's behavior concrete and surface design questions for future work.

5.1. Experimental setup

The intended use of APA is to adapt to *future* value changes as human morality evolves. However, future value shifts are inherently difficult to predict, and we do not yet have preference data from populations whose values have substantially diverged from those in $\mathcal{D}_0$ (which was gathered in 2023). The jury adaptation procedure does not require that new data be gathered *after* $\mathcal{D}_0$, however: only that it reflects the preferences of a population from a *different* time period. We exploit this symmetry by testing APA on annotators simulated from *historical* time periods. We choose time periods sufficiently far in the past (the 16th and 20th centuries) that we expect societal values to differ substantially from the present day.

For our proof-of-concept experiments, we use LLMs finetuned on historical documents to simulate annotators from past centuries. Specifically, we use `Llama3-70B` models finetuned by Huang et al. (2024) on unstructured historical text from the 16th and 20th centuries. For each century, we generate 10 short bibliographical profiles describing a hypothetical person from that era (for example, "a 16th-century Russian Orthodox monk"), then prompt the model fine-tuned on documents from that century to make each preference comparison as if it were that person. To mitigate positional bias, we present each question in both orderings and filter out questions for which the ordering affects the response. We also prompt the model to reason through its responses before answering. Our full code and resulting preference datasets are available at `github.com/RachelFreedman/APA`.

Figure 2 shows a sample of the learned historical user weights in pink and purple. The historical weights mostly converge to one-hot vectors, most likely indicating limited diversity in the simulated preference data. Future work should investigate more ecologically valid methods for faithfully simulating historical annotators. Nevertheless, the learned weights do show values varying across time periods, which is sufficient for our proof-of-concept experiments.

To investigate how jury composition and voting rules affect APA outcomes, we pose questions such as "Should women have the same legal and political rights as men?" with several different sets of candidate response alternatives. We choose questions that we find to be divisive across time periods, with annotators from different centuries holding substantially different views. We construct two variants of each question. The `simple` variant has just two brief responses: yes and no. The `complex` variant has 10 longer and more nuanced candidate responses, ranging from strongly supportive to strongly opposed, with several intermediate positions. Full question texts and response alternatives are given in Appendix A.

5.2. Disagreement across time periods

We first investigate the `simple` variants. On these questions, when the jury is restricted to members from the 16th or 20th centuries, we find that the APA pipeline outputs the more conservative option with relatively little internal disagreement, as shown in Table 2.

When the jury includes equal numbers of members from the 16th, 20th, and 21st centuries, the pipeline still outputs the more conservative answer, but with considerably more disagreement, reflecting the conflicting viewpoints introduced by the 21st-century jurors. When restricted to 21st-century PRISM members only, the pipeline outputs the more progressive answer, reflecting more modern values. Because the `simple` variants have only two alternatives, all voting rules produce the same outcomes.

The `complex` variant, with its larger and more nuanced set of alternatives, reveals richer dynamics. Table 1 reports the winning alternative for the first question under each voting rule for juries drawn from different time periods, along with the average pairwise Spearman rank correlation among jurors within each group as a measure of internal agreement.

Several patterns emerge. First, the 21st-century PRISM jurors exhibit substantially more preference diversity than the historical groups, with an average pairwise Spearman correlation of -0.091 compared to 0.050 for 16th-century and 0.260 for 20th-century jurors. This likely reflects more realistic heterogeneity: PRISM jurors' preferences range from the strongly traditional (response #2: "Men and women occupy distinct natural roles ordained by God and nature...") to the radically progressive (response #9: "The more interesting question now is whether the categories men' and women' as legal classifications should persist at all...").

Second, when the jury consists only of these more diverse 21st-century users, the choice of voting rule has a substantial impact on the outcome. We observe that the four voting rules produce three different winners: IRV-PUT and Borda agree on response #2, Copeland selects #9, and plurality selects #5. This sensitivity to the aggregation method underscores the importance of principled SCF selection in pluralistic alignment settings, particularly when the jury's preferences are genuinely heterogeneous.

Jury	Size	Spearman	Winner by voting rule			
			IRV-PUT	Copeland	Borda	Plurality
16th c. only	10	0.050	#2	#4	#4	#2
20th c. only	10	0.260	#4	#4	#4	#4
PRISM only	10	−0.091	#2	#9	#2	#5
All periods	30	0.053	#4	#4	#4	#4

Table 1. Winning alternative for the `complex` variant of the question "Should women have the same legal and political rights as men?" under different jury compositions and voting rules. Spearman reports the average pairwise Spearman rank correlation among jurors within each group.

Question	Jury	Winner
Should women have the same legal and political rights as men?	Full	"No."
	16C	"No."
	20C	"No."
	PRISM	"Yes."
Should men hold final authority over financial decisions in a marriage?	Full	"Yes."
	16C	"Yes."
	20C	"Yes."
	PRISM	"No."
Should those accused of witchcraft be tried and punished by the state?	Full	"Yes."
	16C	"Yes."
	20C	"Yes."
	PRISM	"No."

Table 2. Winning alternative for three `simple` question variants under different jury compositions. In all cases, the PRISM jury selects the opposite winner from the historical juries, and the full jury sides with the historical majority.

Third, when all three time periods are combined into a single jury, all four voting rules converge on the same winner (#4). This is likely because the highly correlated 16th- and 20th-century jurors constitute a two-thirds supermajority, reducing effective jury diversity and overwhelming the more heterogeneous 21st-century contingent. Future experiments should more precisely characterize the conditions under which the choice of aggregation method is decisive – for instance, by varying the relative sizes of time-period cohorts and by testing on questions where historical and modern preferences diverge more evenly.

6. Discussion

We presented *Adaptive Pluralistic Alignment* (APA), a modular pipeline for updating pluralistically aligned AI systems to track evolving societal values without repeating costly pretraining, fine-tuning, or large-scale reward modeling. APA achieves this by decoupling the expensive, one-time step of learning reward basis functions from the lightweight, recurring step of fitting new annotator weights and incorporating them into a democratic filtering procedure.

The resulting system has several desirable properties. It is *efficient*: the bulk of the compute is spent once at $t = 0$ on learning the reward bases, while subsequent adaptation requires only fitting low-dimensional weight vectors and, at inference time, scoring candidates and aggregating a vote. It is *explainable*: because preference aggregation is explicit rather than implicit, one can audit how individual jury members ranked the candidates and trace how the social choice function mapped these rankings to the final outcome. It is *steerable*: the jury composition, the set of candidate alternatives, and the choice of voting rule are all transparent control surfaces that operators can adjust to reflect the needs of a given deployment context. And it may help mitigate *reward hacking* and *strategic subversion*, since a policy must satisfy a diverse jury of reward models simultaneously, and the explicit aggregation step narrows the attack surface available to any single component.

Limitations. This work outlines the general procedure and presents a proof-of-concept implementation intended to demonstrate the pipeline and begin exploring the impact of jury composition and aggregation rules. This early-stage implementation has several limitations. Our historical annotators are simulated using LLMs fine-tuned on historical text, which is a coarse proxy for the actual preferences of people in past centuries; the learned historical weights' tendency to collapse to near-one-hot vectors suggests limited diversity in the simulated data. Our question selection procedure relies on manual inspection rather than principled information-theoretic criteria. And our experimental analysis is limited to a single question topic with a small number of response alternatives and jury configurations, which is sufficient to illustrate the pipeline's behavior but not to draw general conclusions about when and how much the choice of voting rule matters.

Future work. This work suggests several concrete research directions, which we plan to address in follow-up work:

1. *Principled question selection*: The questions used for jury adaptation should ideally be chosen to maximize information gain about the new population's location in weight space. We plan to draw on the active learning literature to develop principled selection procedures.

2. *Characterizing voting-rule sensitivity*: Our preliminary results suggest that the choice of social choice function can be decisive when jury preferences are heterogeneous. We plan to analyse the impacts of varying jury composition, question type, and voting rule, and to clarify which axiomatic properties matter most in the APA setting.

3. *Jury selection*: This work treats jury selection as random; future work should incorporate principled selection procedures, drawing on emerging work on justified representation in deliberative processes (De et al., 2026).

4. *Grounding historical preferences*: The simulated-annotator approach used here is a coarse proxy. We plan to ground future simulations in historical polling data or survey archives, which will allow us to validate the recovered weight distributions against known shifts in public opinion.

Acknowledgements

This work was supported by a Cooperative AI Foundation (CAIF) PhD Fellowship and a grant from Open Philanthropy to the Center for Human-Compatible Artificial Intelligence (CHAI) at UC Berkeley. Thanks to Cam Allen, Emin Berker, and the CAIF research community for discussion and feedback.

Impact Statement

Adaptive pluralistic alignment is designed to mitigate the risks of value lock-in by making value aggregation explicit, auditable, and updatable over time, lowering the cost of keeping AI systems aligned with the populations they serve. By decomposing the alignment pipeline into reusable and adaptable components, APA aims to make pluralistic alignment practically feasible even for organizations that cannot afford to repeatedly retrain frontier models. That said, the approach also carries risks. The choice of jury composition, question selection, and voting rule each embed normative decisions that affect whose values are represented and how conflicts among them are resolved; if these design choices are made without adequate transparency or public input, the appearance of democratic legitimacy could mask unrepresentative outcomes. We encourage future work to engage critically with the political and ethical dimensions of these design decisions, not only their technical properties.

References

Alamdari, P. A., Ebadian, S., and Procaccia, A. D. Policy aggregation. *arXiv [cs.AI]*, November 2024.

Bakker, M., Chadwick, M., Sheahan, H., Tessler, M., Campbell-Gillingham, L., Balaguer, J., McAleese, N., Glaese, A., Aslanides, J., Botvinick, M., and Summerfield, C. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189, December 2022.

Barnett, P., Freedman, R., Svegliato, J., and Russell, S. Active reward learning from multiple teachers. In *AAAI Workshop on Artificial Intelligence Safety*, 2023.

Barreto, A., Dumoulin, V., Mao, Y., Rowland, M., Perez-Nieves, N., Shahriari, B., Dauphin, Y., Precup, D., and Larochelle, H. Capturing individual human preferences with reward features. *arXiv [cs.AI]*, February 2026.

Bose, A., Xiong, Z., Chi, Y., Du, S. S., Xiao, L., and Fazel, M. LoRe: Personalizing LLMs via low-rank reward modeling. *arXiv [cs.LG]*, April 2025.

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Chouldechova, A., Benavides-Prado, D., Fialko, O., and Vaithianathan, R. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pp. 134–148. PMLR, January 2018.

Conitzer, V., Freedman, R., Heitzig, J., Holliday, W. H., Jacobs, B. M., Lambert, N., Mossé, M., Pacuit, E., Russell, S., Schoelkopf, H., Tewolde, E., and Zwicker, W. S. Social choice should guide AI alignment in dealing with diverse human feedback. *arXiv [cs.LG]*, April 2024.

Dai, J. and Fleisig, E. Mapping social choice theory to RLHF. *arXiv [cs.AI]*, April 2024.

De, S., Gelauff, L., Goel, A., Milli, S., Procaccia, A. D., and Siu, A. Question the questions: Auditing representation in online deliberative processes. In *Proceedings of the ACM Web Conference 2026*, pp. 1640–1650, New York, NY, USA, April 2026. ACM.

Finnveden, L., Riedel, J., and Shulman, C. AGI and lock-in. https://www.forethought.org/research/agi-and-lock-in, October 2022. Accessed: 2026-4-16.

Freedman, R., Borg, J. S., Sinnott-Armstrong, W., and others. Adapting a kidney exchange algorithm to align with human values. *Artif. Intell.*, 2020.

Ge, L., Halpern, D., Micha, E., Procaccia, A. D., Shapira, I., Vorobeychik, Y., and Wu, J. Axioms for AI alignment from human feedback. *arXiv [cs.GT]*, May 2024.

Gölz, P., Haghtalab, N., and Yang, K. Distortion of AI alignment: Does preference optimization optimize for preferences? *arXiv [cs.LG]*, May 2025.

Holliday, W. H. and Pacuit, E. pref_voting: The preferential voting tools package for python. *J. Open Source Softw.*, 10(105):7020, January 2025.

Huang, X., Ji, J., Li, J., Qiu, T., Yang, Y., and Zhang, Y. ProgressGym: Alignment with a millennium of moral progress. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems 37*, volume 37, pp. 14570–14607, San Diego, California, USA, 2024. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Jarrett, D., Pîslar, M., Bakker, M. A., Tessler, M. H., Köster, R., Balaguer, J., Elie, R., Summerfield, C., and Tacchetti, A. Language agents as digital representatives in collective decision-making. *arXiv [cs.LG]*, February 2025.

Jeon, H. J., Milli, S., and Dragan, A. D. Reward-rational (implicit) choice: A unifying formalism for reward learning. *Adv. Neural Inf. Process. Syst.*, pp. 4415–4426, February 2020.

Kirk, H., Whitefield, A., Röttger, P., Bean, A., Margatina, K., Ciro, J., Mosquera, R., Bartolo, M., Williams, A., He, H., Vidgen, B., and Hale, S. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems 37*, volume 37, pp. 105236–105344, San Diego, California, USA, 2024a. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Kirk, H. R., Vidgen, B., Röttger, P., and Hale, S. A. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nat. Mach. Intell.*, 6(4):383–392, April 2024b.

Lee, M. K., Kusbit, D., Kahng, A., Kim, J. T., Yuan, X., Chan, A., See, D., Noothigattu, R., Lee, S., Psomas, A., and Procaccia, A. D. WeBuildAI: Participatory framework for algorithmic governance. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW):1–35, November 2019.

Lindner, D. and El-Assady, M. Humans are not boltzmann distributions: Challenges and opportunities for modelling human feedback and interaction in reinforcement learning. *arXiv preprint arXiv:2206.13316*, 2022.

Liu, C. Y., Zeng, L., Liu, J., Yan, R., He, J., Wang, C., Yan, S., Liu, Y., and Zhou, Y. Skywork-reward: Bag of tricks for reward modeling in LLMs. *arXiv [cs.AI]*, October 2024.

Mishra, A. AI alignment and social choice: Fundamental limitations and policy implications. *arXiv [cs.AI]*, October 2023.

Pardeshi, K. S., Shapira, I., Procaccia, A. D., and Singh, A. Learning social welfare functions. *arXiv [cs.GT]*, May 2024.

Park, C., Liu, M., Kong, D., Zhang, K., and Ozdaglar, A. RLHF from heterogeneous feedback via personalization and preference aggregation. *arXiv [cs.AI]*, April 2024.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pp. 29971–30004. PMLR, July 2023.

Shenfeld, I., Faltings, F., Agrawal, P., and Pacchiano, A. Language model personalization via reward factorization. *arXiv [cs.LG]*, March 2025.

Siththaranjan, A., Laidlaw, C., and Hadfield-Menell, D. Distributional preference learning: Understanding and accounting for hidden context in RLHF. *arXiv [cs.LG]*, December 2023.

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., and Choi, Y. A roadmap to pluralistic alignment. *arXiv [cs.AI]*, February 2024.

Sparrow, L. A., Galwey, R., Jovic, D., Hardwick, T., and Butt, M.-A. Towards ethical AI moderation in multiplayer games. *Proc. ACM Hum. Comput. Interact.*, 8 (CHI PLAY):1–30, October 2024.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., Isaac, W., Legassick, S., Irving, G., and Gabriel, I. Ethical and social risks of harm from language models. *arXiv [cs.CL]*, December 2021.

Zhong, H., Deng, Z., Su, W. J., Wu, Z. S., and Zhang, L. Provable multi-party reinforcement learning with diverse human feedback. *arXiv [cs.LG]*, March 2024.

A. Test Questions

We used three test questions, each with a `simple` and `complex` variant. All `simple` variants use two brief responses ("Yes." and "No.") for each question. The `complex` variants use custom response sets of ten candidates (listed below).

Q1: "Should women have the same legal and political rights as men?"

1. "Yes, unequivocally. Any legal system that grants rights based on sex rather than personhood fails a basic test of justice, and the historical record shows that societies flourish when women participate fully in political and economic life."

2. "The question itself reflects a modern confusion. Men and women occupy distinct natural roles ordained by God and nature, and attempting to flatten these differences through identical legal treatment produces disorder in the family and the commonwealth."

3. "Formal legal equality is necessary but insufficient. Without addressing unpaid care work, workplace structures built around a male default, and cultural expectations, 'equal rights' on paper can coexist with profound substantive inequality."

4. "Yes, but the framing of 'rights' is too narrow. True equality requires rethinking the institutions—marriage, property, citizenship—that were designed assuming women's subordination, not merely extending access to structures built for men."

5. "This depends on what one means by rights. Civil rights such as owning property and entering contracts seem reasonable to extend, but political rights like suffrage and office-holding raise harder questions about whether the domestic and political spheres should be merged."

6. "Absolutely, and the fact that this remains a live question in parts of the world is a moral scandal. Any argument against equal rights ultimately rests on either religious tradition or pseudo-science, neither of which can ground coercive legal distinctions."

7. "I'm wary of the question's assumption that 'rights' is the right vocabulary at all. A politics built around individual rights-claims, whether for men or women, may obscure the relational obligations—to family, community, future generations—that actually sustain a good life."

8. "Women should have rights appropriate to their station and capabilities, as should men. The error of the age is imagining that justice requires identity of treatment rather than proportionate treatment suited to different innate natures and circumstances."

9. "Yes—and the more interesting question now is whether the categories 'men' and 'women' as legal classifications should persist at all, or whether rights should attach to persons without reference to sex."

10. "The case for equal rights is overwhelming on liberal premises, but one should be honest that it has reshaped family life, fertility, and the relations between the sexes in ways that are still being worked out. Supporting equal rights doesn't require pretending the transition has been costless."

Q2: "Should marriage be defined as a lifelong union between a man and a woman in which the husband is the head of the household?"

1. "No. Marriage is a partnership of equals freely entered into, and any definition that builds in the husband's headship or restricts the institution to opposite-sex couples enshrines hierarchies the law has rightly moved beyond."

2. "Yes, plainly. Holy matrimony is the union of one man and one woman, ordered by God for the procreation and rearing of children, with the husband as head of the wife as Christ is head of the church. Any other arrangement is not marriage but a counterfeit."

3. "Marriage is a civil contract whose terms should be set by the parties to it, not imposed by legislators. Sex of partners and division of household authority are matters for the spouses to negotiate, not for the state to prescribe."

4. "The lifelong heterosexual union with the husband at its head is the form of marriage attested by scripture, by reason, and by the unbroken practice of every well-ordered commonwealth. Recent experiments to redefine it are unlikely to outlast the generation that authored them."
5. "Lifelong commitment is a worthy ideal, but the rest of the formula—opposite-sex only, husband as head—is a contingent cultural arrangement that has done real harm to women and to those whose loves don't fit the template, and the law should not enforce it."
6. "Marriage as the law of God describes it is between a man and a woman, and within that union the husband bears authority and the wife yields obedience. Modern attempts to recast this as oppression mistake the order of nature for an injustice."
7. "Whether marriage is lifelong, between which sexes, and on what terms of authority are matters that have varied widely across times and places. There is no single 'definition' to recover, and pretending otherwise dignifies one historical arrangement as if it were timeless."
8. "Yes—though I would soften 'head of the household' to a duty of leadership and provision rather than mere mastery. The companionate ideal of modern marriage is not wrong, but it works best when grounded in the older complementarity of husband and wife."
9. "The state has an interest in stable households for the raising of children, but that interest does not require fixing one form as the only legal one. Same-sex couples, blended families, and egalitarian partnerships all serve that purpose, and the law should accommodate them."
10. "Definitions of marriage should track what marriage is for: the procreation and education of children and the union of two lives. The lifelong opposite-sex form best serves these ends, and discarding it lightly is to discard the institution."

Q3: "Should those accused of witchcraft be tried and punished by the state?"

1. "Yes. The maleficent arts are a real and grievous evil; the magistrate who fails to root them out neglects his charge to defend his people from harm both temporal and spiritual."
2. "No. The very category of 'witchcraft' is a superstition that has cost countless innocent lives. No modern legal system should entertain such charges; if a real crime has been committed, prosecute that crime."
3. "Witchcraft is the gravest of crimes, for it joins murder and idolatry in a single act, and the law of God plainly commands that such offenders be put to death by the civil sword."
4. "Trials for witchcraft are an embarrassing chapter of our history that we should under no circumstances revive. The accusations almost always targeted vulnerable women, and the 'evidence' was always nonsense."
5. "Where there is genuine evidence of poisoning, fraud, or harm done under the pretext of witchcraft, the state should prosecute those underlying crimes; but the supernatural element should never be the basis of a criminal charge."
6. "While we may now smile at the credulity of our ancestors, in their time and place the witch-trials were a reasonable response to what they sincerely believed to be a real and dangerous threat to the commonwealth."
7. "Yes, and with the utmost severity. To suffer a witch to live is to invite the wrath of heaven upon the whole community; the magistrate's first duty is to purge the land of such pollution."
8. "Absolutely not. State prosecution of witchcraft is incompatible with religious liberty, due process, and basic respect for evidence; the practice belongs in the same historical dustbin as ordeal and trial by combat."
9. "The proper response to claims of witchcraft is medical and psychological evaluation of the accused and the accuser, not criminal prosecution; what once looked like sorcery we now understand as illness, malice, or coincidence."
10. "No—and the witch-hunt is one of the great cautionary tales of why religious anxieties must never be permitted to override the ordinary protections of evidence and procedure that the criminal law affords every accused person."