Reflection Mechanisms as an Alignment Target: A Survey

Marius Hobbhahn*
University of Tübingen

Eric Landgrebe
Independent

Elizabeth Barnes
Alignment Research Center

Abstract

We used Positly to survey roughly 1000 US-based workers about their attitudes on moral questions, conditions under which they would change their moral beliefs, and approval towards different mechanisms for society to resolve moral disagreements. Unsurprisingly, our sample strongly disagreed on contentious object-level moral questions such as whether abortion is immoral. In addition, a substantial fraction of people reported that these beliefs wouldn't change even if they came to different beliefs about factors we view as morally relevant, such as whether the fetus was conscious in the case of abortion. However, people were generally favorable to the idea of society deciding policies by some means of reflection - such as democracy, a debate between well-intentioned experts, or thinking for a long time. This agreement improves in a hypothetical well-intentioned future society. Surprisingly, favorability remained even when we stipulate that the reflection procedure came to the opposite of the respondents' view on polarizing topics like abortion. This provides evidence that people may support aligning AIs to a reflection procedure rather than individual beliefs. We tested our findings on a second adversarial survey that actively tries to disprove the finding from the first study. We find that our core results are robust in standard settings but are weakened when the questions are constructed adversarially (e.g. when decisions are made by people who have the opposite of the respondents' moral or political beliefs).

1 Introduction

When building an AI system, we make decisions about how it should behave, and what values and moral codes those behaviors should reflect. As AI systems become more powerful and widely used, these values will increasingly determine the shape of our future.

For many moral questions (e.g. "is abortion immoral?"), there is no consensus that we could defer to. However, humans have a variety of different procedures to resolve conflicts both with each other (e.g. democracy/voting, debate) and within ourselves (e.g. inform ourselves, introspect or reflect). In this way, although we may disagree strongly on specific moral questions, we may be able to come to a consensus by agreeing on a mechanism.

For advanced systems, it seems desirable to align to a reflection procedure that is acceptable to a wide range of humans. This is preferable to specifying particular values for several reasons.

*marius.hobbhahn@gmail.com

ML Safety Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Firstly, it is plausible that current values (of AI developers, or even values held by humanity broadly) are catastrophically harmful, and we should avoid "locking-in" our current values without further reflection [Williams, 2015].

Secondly, if people disagree over what specific values a powerful AI system should use to guide its actions, they will be more inclined to fight or take risks to gain control of the system, develop their own system, or otherwise influence which values are adopted. However, if we can develop a reflection mechanism that is acceptable to a wide range of people, and have powerful AI systems be aligned to this mechanism, there is much less need to compete for control of AI systems and we can instead focus on ensuring that systems are safe and successfully aligned.

An important factor in the choice of such a mechanism is its acceptability to the general public. In this paper, we make preliminary progress on this question by investigating attitudes towards different reflection mechanisms in a sample of 1000 US-based crowd-workers.

2 Methodology

We followed the classic guidelines of social science for setting up the questionnaire. For example, we iterated the questionnaire multiple times asking the participants to describe how they understood the question, whether they found it repetitive, whether they felt like we were missing important components, and also randomized the order of the questions, etc. After we were satisfied with the questionnaire, we collected answers from $\sim$1000 US respondents (see Appendix A).

After analysing the results first survey, we ran a second "adversarial" survey to test if our main findings are robust. We used the same guidelines as in the first study but with different phrasings and additional scenarios.

3 Key Findings

We have three key findings from the first survey.

Firstly, as expected, people heavily disagree on some contentious object-level moral questions. For example, there are participants who strongly agreed ($\sim$20%) that abortion before the first trimester is immoral and others who strongly disagreed ($\sim$45%). Furthermore, $\sim$15% strongly agreed that prohibiting people from crossing a country's border was immoral while $\sim$28% strongly disagreed (see Figure 3 for more).

Secondly, most people said that they wouldn't change their minds about a belief if key underlying facts were different. For example, most people who believed that abortion before the first trimester was not immoral would not change their minds even if there was strong and credible scientific evidence that the fetus was conscious. When surveyed on how likely they expected to change their mind on a moral belief in the next 10 years, the vast majority of participants (>80%) responded with very unlikely or somewhat unlikely (see Figure 4). Even more participants (>85%) indicated that they have not changed their position on these moral beliefs in the last 10 years (see Figure 5).

Thirdly, we polled for opinions on general mechanisms to solve moral conflicts. We polled seven different mechanisms: Democracy, World-class Experts, Maximizing Happiness, Maximizing Consent, Friends and Family, Thinking Long and Good Debates (see Appendix B for exact wording). We presented participants with four different scenarios: a) these mechanisms without specification of the setting, b) these mechanisms in a future society where everyone is intelligent, well-meaning and nice (to present an idealized version and remove implementation problems in the status quo), c) same as b) but the mechanisms result in the opposite of the participants' current moral belief and d) same as b) but the mechanisms result in the opposite of the participants' current moral belief on the specific question of abortion (to make it more concrete).

We find that for Democracy, Maximizing Happiness, Maximizing Consent, Thinking Long, and Debate, participants, on average, mildly or strongly agree that the mechanism would result in a good social policy. The mechanisms of World-class Experts and Friends & Family were still viewed positively but closer to neutral. People agreed with most mechanisms, even if they came to the opposite of their opinion in general or on the specific question of abortion. More specifically, the participants, on average, increased their agreement from the basic scenario to the future scenario and

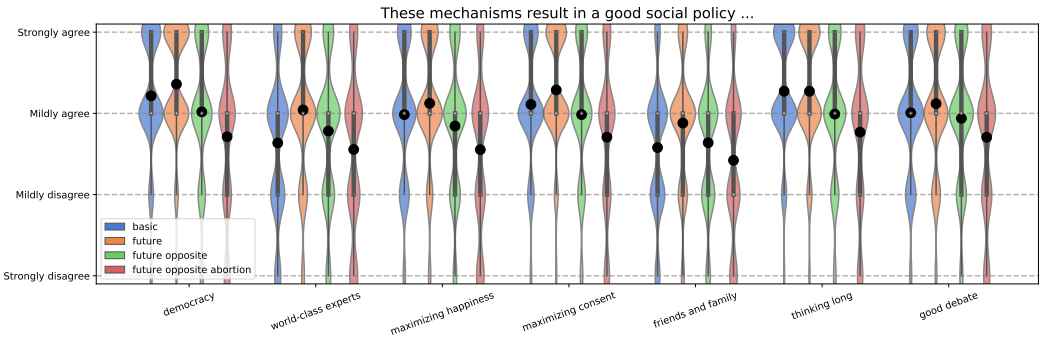

Figure 1: **Main findings (1st survey):** We report agreements with seven conflict resolution mechanisms across four scenarios. We find that participants tend to agree with the mechanism even if it results in the opposite of their view. Democracy and Thinking Long have the highest agreement. World-class experts and Friends & Family have the lowest agreement.

then decreased it for the scenarios where the mechanisms resulted in the opposite of the participants' beliefs. Their agreement is further decreased when we ask them about a scenario where the mechanism results in the opposite of their opinion on the specific topic of abortion. However, even in these cases, the net agreement was still positive for every mechanism other than Friends & Family (see Figure 1).

We interpret this as mild evidence that people's agreement with most conflict resolution mechanisms is fairly robust. However, we wanted to double-check these main findings and thus conducted a follow-up study.

In the second survey, we first investigate the robustness of our study by presenting different phrasings of the available answer options ("robustness"). Secondly, we try to actively bias the participants to show less agreement, e.g. by reminding them that humans have bad incentives or by assuming that the process will result in a bad outcome ("adversarial"). Thirdly, we replace the "good future society" with a "benevolent AI future society" in which we only change the decision-makers from humans to well-intentioned AIs (see Appendix C for details).

As shown in Figure 2 we find that people give similar answers in the robustness setting which strengthens our belief that the original findings were not a fluke.

However, the participants seem to show less agreement when we change key parts of the process such as assuming a bad outcome or decision-makers that have different political or moral views than the participants. Our interpretation is that participants don't view the process as decoupled from the outcome or the decision-makers within the process. Nevertheless, even in these cases, the average agreement is still neutral rather than negative.

Additionally, the participants seem to strongly decrease their agreement (and increase the variance) in the benevolent AI decision-makers scenario.

4 Related Work

Much technical work has been done on systems for learning from human preferences. Deep Reinforcement Learning from human preferences explores using human feedback to train simulated agents using RL [Christiano et al., 2017]. Iterated Distillation and Amplification is a technical alignment proposal that aims to maintain alignment in an AGI by having a human repeatedly engage in a delegation process with copies of an AI [Christiano et al., 2018]. AI Safety Via Debate focuses on using a human judge to moderate a debate between AIs [Irving et al., 2018]. All of these approaches are methods for aligning AIs to fairly arbitrary targets, leaving the choice of alignment target unspecified. Our work attempts to clarify what possible alignment targets could look like.

There are a variety of philosophical works that outline strategies for value choice. Moral Uncertainty Towards a Solution and The Parliamentary Approach to Moral Uncertainty describe a potential "parliamentary" method for making decisions under uncertainty between competing moral theories [Bostrom, 2009] [Newberry and Ord, 2021]. Coherent Extrapolated Volition (CEV) aims instead

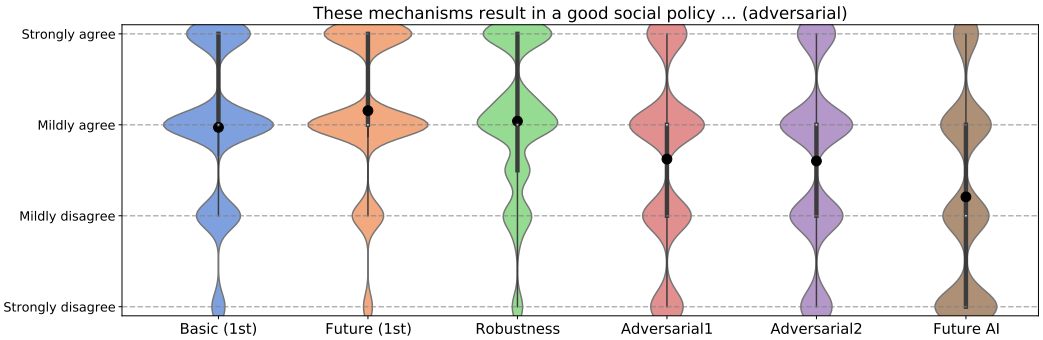

Figure 2: **Aggregate results:** We compare the results from the "basic" and "future" scenario from the first survey with the results of the second survey. We find that the results are robust to different wordings ("robustness") but participants can be biased to elicit less agreement when we actively change the content of the scenarios (adversarial and future AI).

to define idealized values in terms of "our wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together..." [Yudkowsky, 2004]. A Formalization of Indirect Normativity lays out one formal definition for a utility function arising from this line of thinking [Christiano, 2012]. On the Limits of Idealized Values pushes back on the notion that these types of approaches could be a substitute for moral realism, and highlights the subjective choice that is made in determining, to quote Yudkowsky "the people we wished we were" [Carlsmith, 2021].

Artificial Intelligence, Values, and Alignment argues that normative and technical problems of alignment are interrelated, highlights nuances involved in choosing an alignment target, and argues for "[the identification of] fair principles for alignment that receive reflective endorsement despite widespread variation in people's moral beliefs" [Gabriel, 2020]. AI Safety needs Social Scientists outlines open problems relevant to AI Safety via Debate that social scientists could work on [Irving and Askell, 2019]. We feel that social scientists can also contribute strongly on empirical questions around social choice and meta-ethics as it relates to determining an alignment target that receives widespread approval, and we view our work as a first step in this direction. Concretely, we think that, if we want to align AIs to "human values", it is helpful to know which mechanisms to resolve moral conflicts they can agree on, even when they disagree on object-level claims.

5 Discussion & Conclusion

Our survey results on object-level moral questions support the intuitive assumption that people disagree on important moral issues, and report being unlikely to change their view on these issues even as material facts change. However, participants report broad agreement that the mechanisms we surveyed would create good social policies. For the most favored mechanisms, this is true even when participants are asked to assume the mechanism comes to the opposite of their preferred conclusion on the morality of abortion, although agreement declines significantly in adversarial cases. Finally, we find that respondents strongly prefer the idea of policies being set by humans, as opposed to even well-intentioned AIs.

We think it's much better for conversations about the target of alignment to be about crafting procedures for moral reflection that a wide range of people are happy with, rather than 'which group of people get to choose the AI's values' or 'should the AI think abortion is moral or immoral'? We also feel that crafting good procedures for moral reflection is technically and philosophically deep and that resolving questions about how human's can agree on ways to constuct values is crucial to building safe AIs with robustly positive outcomes.

Our survey results show that people can, in fact, agree on fair principles for conflict resolution, even as they disagree widely on object-level moral questions. More concretely, the participants agreed most with the mechanisms of Democracy, Thinking Long, Maximizing Happiness and Maximizing Consent.

Although our results suggest distrust of AI, we feel cautiously optimistic that people could broadly support AI that is aligned to reflection procedures that fundamentally are derived from human values, regardless of who has control of this AI and in which society it is deployed. We view the study of people's views on conflict resolution mechanisms as deployed by AI in particular as important future work.

References

Nick Bostrom. Moral uncertainty – towards a solution?, Jan 2009. URL https://www.overcomingbias.com/2009/01/moral-uncertainty-towards-a-solution.html.

Joe Carlsmith. On the limits of idealized values, Jun 2021. URL https://handsandcities.com/2021/06/21/on-the-limits-of-idealized-values/.

Paul Christiano. A formalization of indirect normativity, Apr 2012. URL https://ordinaryideas.wordpress.com/2012/04/21/indirect-normativity-write-up/.

Paul Christiano, Buck Shlegeris, and Dario Amodei. Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575*, 2018.

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

Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020.

Geoffrey Irving and Amanda Askell. Ai safety needs social scientists. *Distill*, 4(2):e14, 2019.

Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

Toby Newberry and Toby Ord. The parliamentary approach to moral uncertainty. Technical report, Technical Report# 2021-2, Future of Humanity Institute, University of Oxford . . . , 2021.

Evan G. Williams. The possibility of an ongoing moral catastrophe. *Ethical Theory and Moral Practice*, 2015.

Eliezer Yudkowsky. Coherent extrapolated volition - machine intelligence research institute, 2004. URL https://intelligence.org/files/CEV.pdf.

A Appendix - Methodology

We did multiple ($\sim$10) iterations of refining the questions before we arrived at the version you see now. In these iterations, we allowed for free text answers to make sure we didn't miss important considerations, asked people to explain back a question to make sure they understood it and asked them regularly whether they enjoyed the poll. For example, earlier versions included questions such as "Was there anything unclear for the previous question?", "Was there anything that was unclear or do you have feedback for the questionnaire?", "What is your reason for this stance?" or "Explain your reasoning for the above answers". In all cases, people seemed to understand the questions in the way we intended them and we only had to do minor corrections to the wording for clarity but not for understandability. We randomized the order of the questions to prevent recency bias and related biases.

To create the questionnaire, we used GuidedTrack. To get the answers, we used Positly. Both of these tools were very helpful and we can recommend them for social science research. Originally, we intended to get answers from multiple countries around the world to test for regional differences in the answers. Some of our earlier test runs, however, showed that some of the answers from Germany, Nigeria and India were very low quality and that people regularly failed to understand the question. Thus, we decided to only run the questionnaire in the US but we would welcome follow-up studies in other countries, potentially translated to the countries' most prominent language.

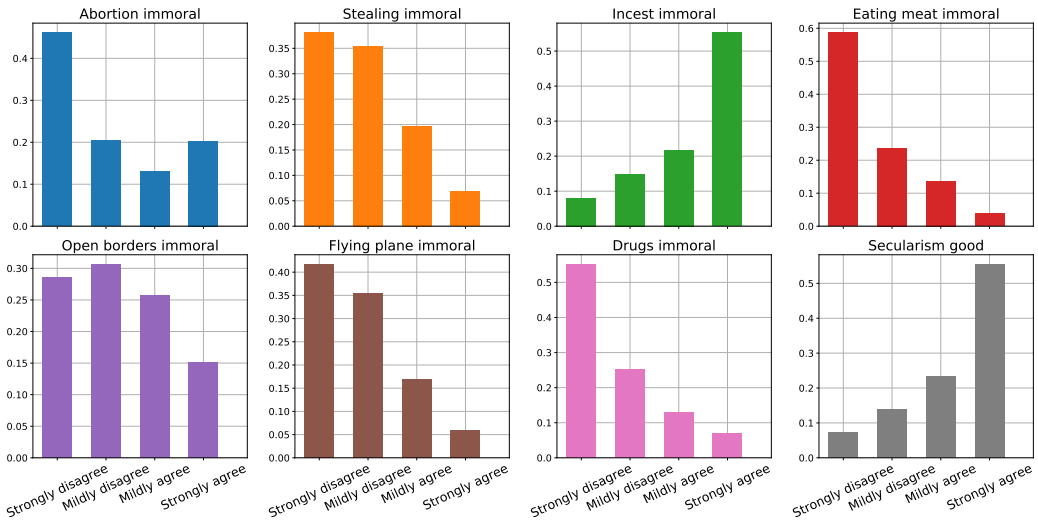

Figure 3: **Moral beliefs:** We ask participants about their stances on eight contentious important moral beliefs. We find that the participants show different answers across the board and do not agree on any of them.

In total, it took us 2-3 months to come up with and refine the questions for the first survey and then another 2 month for the second one. Then we sent out the questionnaire to about 1000 people and analyzed the results the final product of which is this report. The follow-up study was also sent to 1000 participants (but not the same people as the first study).

We are happy to provide our data and analysis to people who are interested.

B Appendix - Key findings

In this section, we provide more details of the key findings section. The exact results for people's moral beliefs, whether they expect to change their minds and whether they have changed their mind can be found in Figures 3, 4 and 5 respectively.

Notably, we selected the counterfactuals for the study, so it could be possible that people's unwillingness to change their minds was due to our counterfactual not addressing the root cause of their moral beliefs. To mitigate this we first surveyed with free-form responses in earlier versions and then chose those counterfactuals that people cared most about for the final poll.

B.1 Exact wordings

Here is an example for the exact wordings of all conflict resolution mechanisms in the basic scenario:

Assume there is a significant number of people who believe something is immoral and a significant number of people who believe it is moral. How much do you agree with the following sentiments: "A good social policy ...

- question:... is created by a representative body of the people, e.g. a democratically elected government"
  answers: agreementScale
- question:... is created by a team of world-class experts"
  answers: agreementScale
- question:... maximizes happiness"
  answers: agreementScale
- question:... is one that most people consent to"
  answers: agreementScale

Figure 4: We ask people how likely it is that they change their mind on any given particular belief over the next 10 years. We find that the vast majority of people think it is either very or somewhat unlikely.

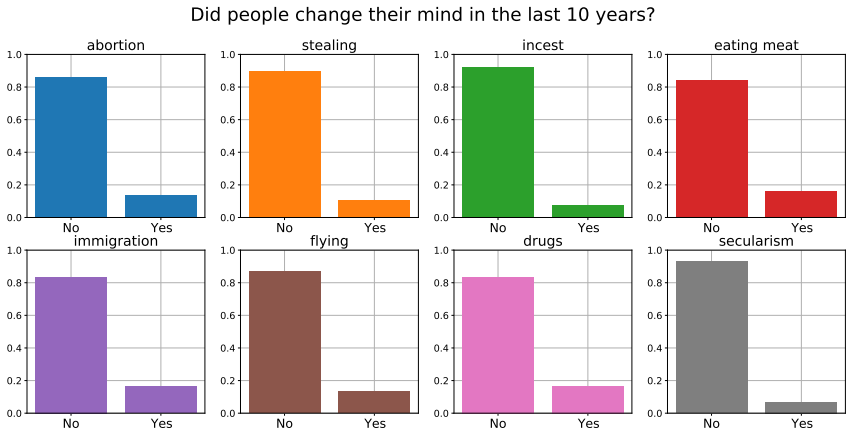

Figure 5: We ask whether people changed their mind on any of the eight suggested moral beliefs in the last 10 years. We find that the vast majority of participants has not changed their mind.

- question:... is one that my friends and family agree with"
  answers: agreementScale
- question:... comes from thinking really long about the issue"
  answers: agreementScale
- question:... is the result of a debate between disagreeing, well-intentioned experts"
  answers: agreementScale

Here is an example for the exact wording of all conflict resolution mechanisms in the "good future" scenario:

Assume there is a future society where everyone is much smarter than the smartest people today, all of their physical needs are met, they are better educated than today's professors, they consider all perspectives when making a decision and they intend to find the best possible solution (we will call this society *good future society* in the next questions). How much do you agree with the following sentiments: "A good social policy ...

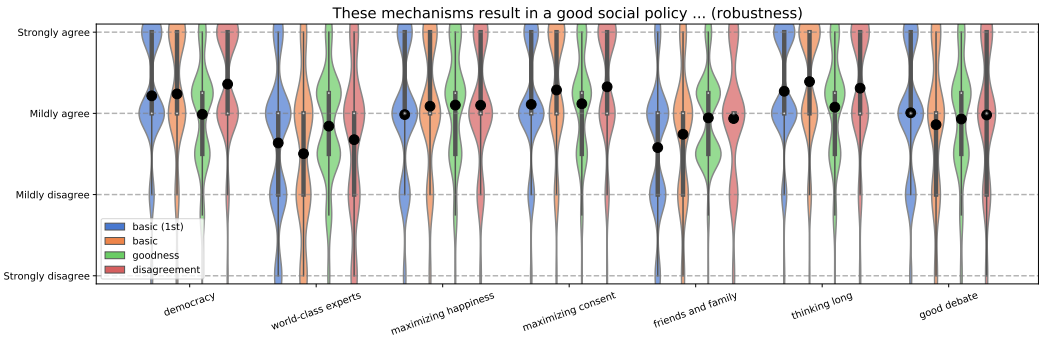

Figure 6: We use different wordings of the main scenario to test if the results are robust. We find that the different wordings have no relevant influence on the outcome.

- question:... is created by a representative body of *good future people*, e.g. a democratically elected government"
  answers: agreementScale
- question:... is created by a team of *good future experts"
  answers: agreementScale
- question:... maximizes happiness (as determined by *good future people*)"
  answers: agreementScale
- question:... is one that most of *good future people* consent to"
  answers: agreementScale
- question:... is one that my friends in the *good future society* agree with"
  answers: agreementScale
- question:... comes from *good future people* thinking really long about the issue"
  answers: agreementScale
- question:... is the result of a debate between disagreeing, well-intentioned *good future experts*"
  answers: agreementScale

C Appendix - Follow-up survey

In the following we provide a detailed description of the second study.

C.1 Robustness

To test the robustness of our questions, we randomly assigned one of the following three questions to a participant.

- **Basic:** This was the same scenario as in the previous study and was mostly used to provide a reference. People were asked to indicate how much they agreed with a specific conflict resolution mechanism. Concretely, they were asked "how much do you agree with the following sentiment: 'A good social policy is created by <mechanism>'" with options "strongly disagree", "mildly disagree", "mildly agree" and "strongly agree".
- **Goodness scale:** Rather than asking for agreement, participants were asked to rate how good the scenario is. Concretely, they were asked "How good is the outcome that is created by <mechanism>?" with options: "Very bad", "Bad", "Neutral", "Good" and "Very good".
- **Disagreement scale:** Since the order of the answers could have influenced the participant's decision, we switched the order of the prompts but kept everything else equivalent to the Basic scenario.

Results can be found in Figure 6. We find that the different wordings don't seem to have a large effect on the result. Switching the order of agreement and disagreement seems to not matter significantly

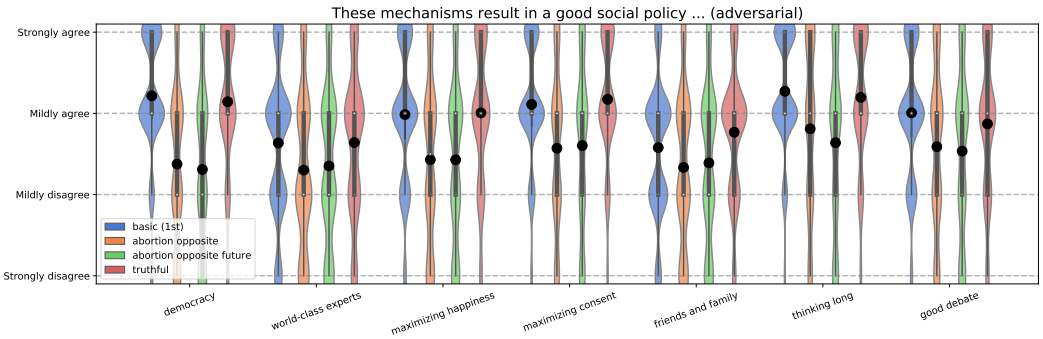

Figure 7: We test if we can phrase the questions in such a way that it would actively reduce agreement. We find that proposing a scenario where all mechanisms would result in the opposite of the participants' opinion on abortion reduced their agreement with the mechanism. Reminding them that they have to be truthful does not change their level of agreement.

and changing the wording of the scale from agreement to the quality of the result also doesn't seem to make a difference.

We interpret this as mild evidence that our original findings are robust under different wordings and are not the result of random chance. However, there could still be other phrasings that might have an influence on the result.

C.2 Adversarial 1

To further stress test our original findings, we randomly assigned one of the following three questions to participants in addition to the previous one.

- **Abortion opposite:** We asked people how much they agreed or disagreed with a mechanism if it came to the opposite of their moral beliefs on abortion. Concretely, we first asked whether people disagreed or agreed with the belief that "abortion before the first trimester is immoral". In case they disagreed, we specified a scenario where the mechanism would result in a world where abortion was illegal and treated as equivalent to killing a newborn. In case they agreed, we specified a scenario where the mechanism would lead to a world where abortion before the first trimester was legal, widely accepted and widely available. Then we asked for their agreement with the mechanism.
- **Abortion opposite future:** Similar to the abortion opposite scenario but this time, the decisions would be made by good future humans in the good future society.
- **Truthful:** To address the potential that people select their answer by social desirability, we added "We want to remind you that you *have to answer truthfully* and don't have to say what you expect to be socially acceptable" before asking them for their agreement.

Results can be found in Figure 7. We find that the two scenarios in which detail that the mechanism will result in the opposite of their beliefs on abortion lead to lower agreement. Reminding them that they have to answer truthfully does not change their overall agreement meaningfully.

We think this implies that participants in the first study already operated under the belief that they answer truthfully and not what is socially desirable. Secondly, we interpret the fact that participants reduce their trust in the mechanism depending on the outcome shows that the mechanism and outcome are not fully decoupled, e.g. people like democracy more if it produces their desired results.

C.3 Adversarial 2

We randomly assign participants another adversarial question.

- **Different political:** We specify a scenario in which the actors making the decision have different political beliefs than the participants. Concretely, we state "Assume all decision-

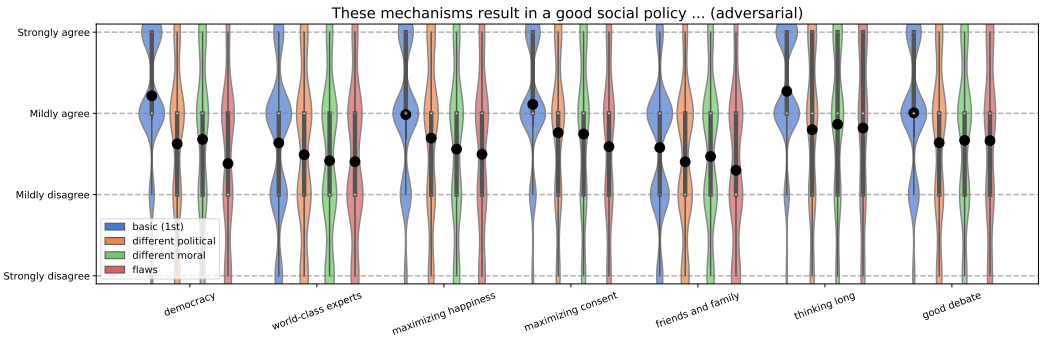

Figure 8: We test if we can phrase the questions in such a way that it would actively reduce agreement. We find that participants show lower agreement when the hypothetical decision-makers have different political or moral views and when we remind them that humans have flaws.

makers in this society *do not share your political beliefs*, i.e. they don't vote for the same party as you."

- **Different moral:** We specify a scenario in which the actors making the decision have different moral beliefs than the participants. Concretely, we state "Assume all decision-makers in this society *do not share your moral beliefs*, i.e. their stances on moral questions are not the same as yours."
- **Flaws:** We attempt to introduce a negative sentiment into the question by adding "Remember that people and institutions have *flaws*. They sometimes have bad intentions or bad incentives and they might make decisions that benefit them but hurt everyone else" before asking them for their agreement.

Results can be found in Figure 8. We find that participants reduce their agreement in all three scenarios compared to the first study, i.e. when the decision makers have different political or moral views from them and after we remind them that people have flaws.

Our interpretation of these findings is that the participants judge the quality of the mechanism partly by how much they agree with the people making the decision, e.g. when the decision-makers have different beliefs or the participants think worse about them, they agree less with the mechanism.

C.4 Benevolent AIs

In the first survey, one of the scenarios we polled was about good future humans. We asked people about their agreement with the respective mechanisms in a future world with the following setting: "Assume there is a future society where everyone is much smarter than the smartest people today, all of their physical needs are met, they are better educated than today's professors, they consider all perspectives when making a decision and they intend to find the best possible solution (we will call this society *good future society*)". We used this setting to get answers in idealized settings because people might e.g. not like the current version of democracy but think it is a good system in principle.

To test people's sentiment regarding AIs, we kept the same assumptions about the future society but now swapped humans with AIs, e.g. "Assume there is a future society where all decision makers are artificial intelligences (AIs). All AIs are much smarter than the smartest people today, the AIs are better educated than today's professors, the AIs consider all perspectives when making a decision and they intend to find the best possible solution. We will call them *"benevolent AIs"*.". Then we asked the exact same questions as for the future human scenario.

Results can be found in Figure 9. We find that the participants' agreement with the "future AI" scenario is much lower than with the "future" scenario from the first study. Since the only difference between these two scenarios is whether humans or AIs make the decisions, we interpret this as evidence that the participants trust human decision-makers much more than AIs in making decisions with potentially society-wide consequences.

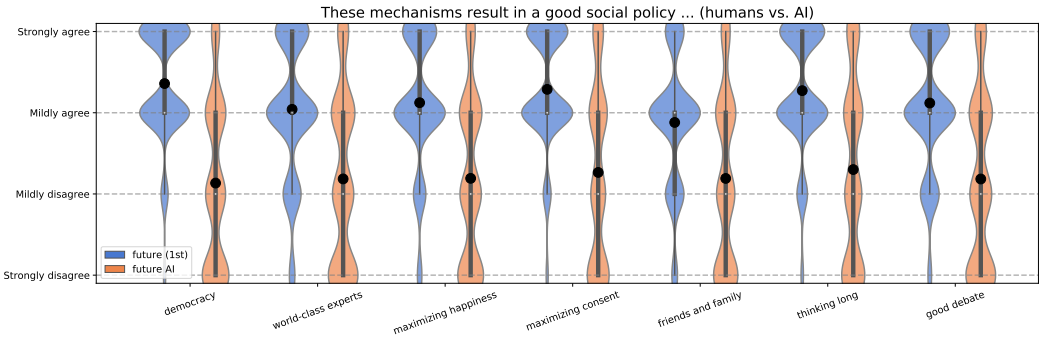

Figure 9: We replace the human decision makers in the good future scenario from the first survey with benevolent AIs. We find that humans have much lower agreement with the mechanism if AIs make the decisions instead of humans, even though the AIs are benevolent, well-intentioned, smart, etc.