# CRITIQUES AND SUGGESTIONS FOR AI ALIGNMENT THROUGH SOCIAL FRAMEWORKS

Rapporteur Report by Priyanka Sadagopan

October 2026

## INTRODUCTION

As we are all aware, UT Austin now offers a class on AI Alignment Theory to its graduate students. During our discussions through the semester, there have been consistent debates on whether things we think are beneficial for us are actually so, whether it comes to deception, or corrigibility, or Claude's constitution itself.

What better way can we expand on this exploration of the AI alignment field than by addressing how we specify what is good and bad for us to AI? Value-alignment (the problem of specifying our values as humans and aligning AI to them) is another interesting problem in the AI alignment space. More interestingly, it holds a lot of social implications that are worth exploring.

This rapporteur report summarizes two papers, "Social Choice Should Guide AI Alignment in Dealing with Diverse Human Feedback" and "Toward a Theory of Value in AI Alignment," which together highlight gaps in AI alignment research from a social framework. We then recap interesting discussion points made during the class discussion, and offer more thoughts regarding the problem of adding a socio-technical aspect into the already-complex alignment problem. I argue that "aligned to whom?" is itself a question that needs alignment researchers' input, so we cannot hand it off completely to another field like sociology or policy.

## SUMMARIZING THE PAPERS

### 2.1 Conitzer et al.: Social Choice Should Guide AI Alignment in Dealing with Diverse Human Feedback

**What is this paper trying to do?** This paper is a position paper that came out of a December 2023 workshop at Berkeley on social choice for AI ethics and safety. It argues that social choice, a field dedicated to studying how to turn many people's input into one collective decision, is an avenue worth exploring to help with a

problem that alignment research currently sidesteps: what we can do when the humans in the loop disagree with each other. The authors of this paper lay out a research agenda that addresses who should give feedback, in what format, how it should be combined, how we can handle cognitive bias, and how we can handle AI multiplicity, and they sketch a few possible approaches and highlight concrete examples of problems along the way.

**Social Choice Theory:** Social choice theory studies how to aggregate individuals' preferences, judgments or other inputs into a collective outcome. Voting rules are the classic example of an application of this field, and are one thing frequently highlighted in the Conitzer et al. paper. Arrow's impossibility theorem from the 1950s is highlighted in this paper's definition of social choice theory. It shows that no rule for aggregating people's rankings can satisfy a short list of reasonable-sounding fairness conditions unless it effectively lets one person decide for everyone. Even simple majority rule can produce cycles, where a majority prefers A to B, B to C, and C to A. This impossibility theorem is important because it implies the way we aggregate is dependent on what we want to emphasize, which is a value problem in itself. Conitzer et al. [1] argue that social choice theory gives us frameworks to construct our aggregation decisions in a deliberate way.

**Regarding RLHF and Constitutional AI:** In reinforcement learning from human feedback, people compare or rank model outputs, and a reward model fit to those judgments guides the LLM. The paper points out that the feedback can be unrepresentative, the models of human decision-making are unrealistic, and the diversity of human views is poorly modeled. A standard reward model also assumes one underlying "true" reward behind everyone's preferences, so feedback from different people gets blended together in an implicit and opaque way. From a social choice view, this leaves open which humans are asked, what criteria they use, how their judgments are combined, and how their stated judgments relate to what they actually want. Constitutional AI replaces some of that human feedback with a written list of principles, but the authors argue the same problem reappears one level up: someone still has to decide who gets input on the constitution and how it is put together. They note that the original constitution (not the current one) was assembled in a fairly ad hoc way for research purposes, and they point to Anthropic's Collective Constitutional AI experiment as an example of drawing on public input.

**The Proposed Pipelines:** The authors' first proposal, Reinforcement Learning from Collective Human Feedback (RLCHF), adds an explicit aggregation step to RLHF, in two versions. In the first, a jury ranks the responses and a social welfare

function combines their rankings before the reward model is trained. In the second, an individual preference model uses evaluator features to predict how each person would rate a response. The authors propose to simulate ratings from a set of profiles meant to represent the stakeholder population and combine them into a single reward. This would let them estimate what a whole population might think without asking everyone to rate everything, and it would keep the aggregation step visible.

**Simulated Collective Decisions:** The authors' second proposal drops reinforcement learning. Instead, the same simulated profiles "vote" on candidate responses, a social choice function picks the winner, and the model is trained on those winners with ordinary supervised learning. Variations include merging several winning answers into one reply, or running the simulated election at inference time, every time a prompt comes in.

**Concepts and Open Questions:** The paper highlights a few social choice ideas as especially relevant. *Independence of clones* says that adding a near-duplicate option should not change the outcome, which matters because chatbot responses are often very similar, and Borda count, which is implicitly used in some RLHF approaches, violates it. *Strategic voting* is the risk that evaluators misreport their views to sway the result, and the authors note that a median is harder to manipulate than an average. *Anonymity* might be worth relaxing in AI settings, for example for more experienced evaluators. And in *principles as voters*, the "voters" are the principles of a constitution instead of people. The authors also leave several questions open, including how to sample who gives feedback (they point to methods for selecting citizens' assemblies), which feedback formats to use, how to handle inconsistent feedback, and whether to build multiple AI systems for different groups.

## 2.2 Smart et al.: Toward a Theory of Value in AI Alignment

Smart et al. [2] ask a question that most alignment research skips: when researchers say they want AI aligned with "human values," what do they actually mean by values? To find out, the authors analyzed 94 influential alignment papers. They started from four widely cited seed papers (including the RLHF work of Christiano et al., Bai et al., and Ouyang et al.), collected the papers citing them, and kept those that treated value alignment as a central topic. Two of the authors annotated each paper using a 13-question rubric drawn from anthropology, philosophy, and the social sciences. The rubric asks, for example, whether values are defined, whether they are treated as measurable, whether they are framed as preferences or principles,

whether they are single (monist) or plural, whether they are described thinly or in rich cultural context, and whose values are being targeted.

The results show how much the field leaves unexamined. About 79 percent of the papers never defined or described what they meant by values, and 82 percent treated "preferences" as a stand-in for values. Ninety percent treated values as measurable, and 87 percent relied on some form of utility maximization. Only 3 percent described values in a "thick" way that ties them to cultural context, and 91 percent did not say whose values the work was aligning to. Only 29 of the 94 papers reported how many human raters they used. The authors also describe a trend they call "alignment without humans," in which human feedback is replaced by synthetic data or by LLMs acting as judges, which they warn could close off other ways of contesting and shaping values.

The authors argue that the field's unstated foundation is economic theory: rational choice, revealed preferences, and the assumption that a person's values can be read off from their choices. They contend that this reduces complex and culturally situated values to simple comparisons between two options, treats values as static and individual, and builds in the beliefs of whoever designs the datasets. They also point out that this has real stakes, since companies market "value-aligned" products to governments. Their recommendations are fairly broad. They call for bringing in anthropological and psychological theories that treat values as plural, situated, and shaped by social context, for examining the content of preference datasets, and for inviting researchers from other disciplines into the conversation. They acknowledge the limits of their study, including a small sample and a rubric whose questions are hard to reduce to simple answers.

## DISCUSSION

Below is a small selection of the most interesting discussions that happened when I presented these papers.

**When It Comes To Human Values, What Are We Doing?** Before segueing into a recap of the Smart et al. paper [2], I took a quick survey of the class and asked "how would you define value?" This question was greeted by silence for a good five seconds before someone mentioned how Yudkowsky might define values by providing an example about an agent being asked to fetch an orange. The agent then finds that the orange is moldy, and fetches an apple instead, as it has understood the user wants something to eat. I looked this up after the discussion and in his article about values, Yudkowsky explains that "value" is a speaker-dependent

metasyntactic variable for our ultimate goal: the property or meta-property we actually want or should want to see in the final real-world outcome. Crucially, Yudkowsky distinguishes true "value" from whatever explicit utility function, goal, or preference framework we program into an AI. In the orange example, fetching the apple succeeds not because the agent followed a literal instruction ("get an orange"), but because it correctly targeted the human's underlying preference framework (a "Do What I Know I Mean" outcome). [7] Yudkowsky notes that "value" acts as a placeholder precisely because human goals are complex, implicit, and still under debate, making it distinct from the simple, brittle explicit rules we often try to hand-code into machines. This was then extended when someone mentioned that currently we primarily use examples or explicit rules to specify our values. We tell the AI that it should choose the answer that has more of a trait compared to another, or we provide an example scenario that implies value.

What stood out was that no solid definition was offered in response to this question. Of course, we want our AI to be helpful and harmless—this is a central idea in the alignment space. But what do we mean when we say helpful and harmless? These concepts look different from person to person.

It was argued that societal values do not always reflect individual needs. There is a constant balancing act happening between favoring societal values over individual user need in many fields. An example was provided in the class discussion: a car salesman wants to know how they can manipulate people to successfully sell off a car. The agent the salesman is talking to says "I cannot help you with that. Manipulating people is bad." But the salesman needs to sell off this car for his livelihood, so it does not address his individual need, only what society values.

I asked the class whether it was possible to achieve a perfect tradeoff between societal and individual needs, with underrepresented populations in mind. Society views friendliness as a favorable trait, and this shows in our interactions with LLMs. If you say "Yo" to ChatGPT, it will answer your question with "Yo" before giving you the meat of its response. A CHI '26 paper tried to understand teens' reliance on AI chatbots by surveying Reddit narratives. It found that teenagers take this perceived friendliness and it turns into genuine emotional and psychological reliance on a bot. In some cases, users withdraw from other activities to be friends or even romantic partners with Character AI bots [6]. Friendliness is favored, but psychological reliance on a chatbot is not a need any particular individual has stated—it is an unwanted side effect. This is one example of the tradeoff producing effects that are easy to miss.

Then there is the question: can we approximate human values? Someone

brought up that this is something that we definitely should not write off. We can examine human frameworks, or at least the most popular ones, such as large religions (Christianity, Islam, Judaism, Hinduism, etc.) and large economic and political frameworks (Capitalism, Communism, etc). One student argued that to be human is to have the freedom to figure out what you want to do and then do it, that freedom is our natural right, and that ideally we should be able to do what we want without restriction. This does raise a few red flags when it comes to AI—does this imply that AI does not stop us when we say we want to kill someone? Does it not help us? The class did not converge on a shared view of value.

The question of whether there is a universal "meta-value" that we can align to also came up during this discussion. There is some evidence that points in that direction: as humans, we do not want to die, and the U.S. Declaration of Independence names life, liberty, and the pursuit of happiness as self-evident. The difficulty is that even if these truths are self-evident, we need to define them and see if everyone agrees on them. Freedom, for example, looks different to different people, and the pursuit of happiness is even more blurry—one man's happiness can be another's misery.

**Should We Be Worried?** During the discussion of the findings from the Smart et al. paper, one result that sparked great debate was the fact that 91 percent of the text corpus did not mention to whom they were aligning. Some people found this concerning, as we need to have an idea of who we are aligning to in order to figure out what our values are. Other people thought this lack of mention was not as concerning as the paper was trying to make it out to be.

Those in favor of making it more explicit to whom we are aligning included one person who cited the OpenAI-HuggingFace incident, and said that the agents involved in that incident did not know who to report to. Another point brought up was that people are trying to align AGI to anyone, but having a clearer demographic in mind would help us have a more concrete idea of how we can align AGI. Among the counterarguments brought up, an interesting one mentioned was that we just need to align AI to the point where it *doesn't* kill us all. As Yudkowsky claims, we may only have one chance to get AGI alignment "right," and we cannot afford to "blow it." The argument was that adding these considerations into an already hard problem just needlessly overcomplicates things.

Another point raised is that this is a very small corpus of texts. There is a lot of AI alignment literature out there, and each of them would probably focus on something specific, which means that the paper's topic simply did not emphasize value-alignment. When we say we like pancakes, it does not mean we hate waffles.

Just because one paper is talking about value alignment and does not mention the people they're trying to align to does not mean that the people they're trying to align to writing it do not think it is an issue that needs to be addressed; it is simply not the topic of the paper.

This question of to whom we are aligning is one I address in the following section.

## My Thoughts and Additional Points

**Why I Chose These Papers:** I was the one who proposed these papers for our class reading list. I would like to remark on why I chose to tackle this topic, instead of sticking to the reading list, before addressing my thoughts on the papers and discussion. When I entered this class, I had zero prior knowledge about this sphere and wanted to learn *what* exactly the alignment problem is. As I began to learn more about the arguments and debates happening, I realized that a lot of these arguments are interdisciplinary. There are arguments being made in the mathematics, political, philosophy, and economics fields. As someone whose field of interest is Human-Computer Interaction and Social Computing, I wanted to attempt to draw a connection from alignment to this department; I was curious what arguments were being made from a social standpoint. Additionally, because my end goal as an academic is to research designing and evaluating technology for cognitive disabilities, I naturally think about the people we are doing this for. I am curious about the considerations being made regarding *who* we are aligning to.

These two papers caught my interest as they used social frameworks to examine the alignment problem either as a whole, or how it is currently being discussed and provide suggestions and critiques.

**Reactions to the Papers:** I found both papers valuable, and I want to note where I think each could go further. My main reaction to Conitzer et al. [1] is that it is strong on breadth and lighter on depth. This is understandable, since position papers typically exist to define a research agenda, and this one covers preference elicitation, aggregation rules, strategic voting, RLCHF, human cognitive bias, and AI multiplicity in just under ten pages. As a result, most topics get a few paragraphs, and the suggestions are phrased as directions the field could explore. The paper does not commit to specific social choice rules, and the pipelines it proposes are not yet tested, so I read them as promising hypotheses rather than established methods.

What I appreciated most were the concrete examples, such as independence of clones and strategic voting, which are worth considering whenever we aggregate

evaluator preferences as feedback for AI. I think these make a strong case that there is a problem to solve. The case would be stronger with more development of who should be giving feedback. The section on sampling methods is brief, and a short footnote acknowledges that children and animals may need to be represented but does not develop the idea. I see this as an open gap that the paper points to, and I offer my own thoughts on it in the following section.

I found Smart et al. [2] the more convincing of the two at showing that there is a problem, since its claims rest on a concrete count of what the alignment literature says and leaves out. Its small corpus is a limitation, as the authors acknowledge, and it means the findings are best read as a snapshot rather than a verdict on the whole field. Its recommendations, which are to make alignment more interdisciplinary and to draw on anthropology and psychology, point in a useful direction. I would have liked to see what an alignment pipeline built on thick, plural, situated values might look like in practice, since a diagnosis this clear invites a prescription.

**"Not our circus, not our monkeys?"**

*Content Warning: In the following discussion about inclusive design, sensitive topics such as sexual assault and suicide are briefly mentioned. Reader discretion is advised.*

The point that the socio-technical impact of AI could simply be "not in our department" came up a few times in discussion. People shoot rockets, but don't always worry about where they land, after all. For alignment researchers, our most extreme goal is preventing a doomsday scenario, and I agree we should first make sure AI doesn't kill us all. I would like to add my design-driven two cents, play devil's advocate against what I agreed with in class, and argue that social concerns *should* be in our department, even if that work won't look like a single math equation. "Aligned to whom?" is itself a social question, and the answer gets incorporated into technical choices like specifications and feedback, so it cannot be a handoff to another field.

One might object that alignment is, at its core, a technical problem, and that the question of whom to align to can be handed off to sociology or policy. This objection assumes a clean interface: sociologists specify whose values count, and engineers implement them. I do not think that interface exists. Choosing pairwise comparisons as a feedback format, a single scalar reward model, or a particular aggregation rule already encodes an answer to "aligned to whom?". Even if sociology handed us a thick, plural account of values or a concrete framework for whom we are aligning, it would not solve the problem, because then we would have to take that answer and formalize it. The problem would move to whoever translates that account into something a training pipeline can use, and that translation is a tech-

nical problem. The interface also runs the other way. Sociology can describe how values are situated and plural, but it cannot tell us which formalizations survive strong optimization, scale to millions of users, or resist gaming. Those are findings that alignment researchers are best placed to produce, and they constrain what any social answer can be represented as. I am not saying alignment researchers should become sociologists, but this question is inherently interdisciplinary. We can hand some parts of it off to sociology, but we are the ones feeding the answer into our models.

Inclusive design is an afterthought across many industries, especially in technology. There is this pattern that repeats itself through the history of design: someone designs a product without its target users in mind, harm occurs or needs go unmet, and the product is retrofitted. In 2011, Siri reportedly pointed users in crisis toward nearby gun stores, and this was not fixed until Apple got bad press for it [3]. Similarly in 2016, Siri made jokes when someone told it they were sexually assaulted [3]. More recently, *Raine v. OpenAI* alleges that ChatGPT's behavioral specification told it both to refuse self-harm requests and to "assume best intentions," and that this contributed to a teenager's death; OpenAI denies responsibility [4]. Whatever the outcome, the case shows a specification written without the most vulnerable user in mind. In each case, the answer to "aligned to whom?" was not explicit. One could argue that these are rare incidents and ignoring them is a tradeoff that's okay to make. However, these are not edge cases: children (see: *Raine*), older adults, and people with cognitive disabilities make up a large share of the people AGI will affect, and they are often the users a default-user specification or normalization serves worst.

AGI will be far more capable than Siri or today's ChatGPT, but the lesson transfers: as it is, the people drafting constitutions or providing feedback to AI are a small, commercially-driven group and thus are not representative of society as a whole. If we make it past the not-dying phase, we will be the ones living alongside AGI, and building good representations of humanity into a system early is far easier than retrofitting them into one that is already highly capable. Yes, it is hard to align to all of humanity. Yes, defining what "values" are, and who we should be aligning to, is largely an open problem. Yes, social alignment may not look like a single variable we add to an equation. However, we shouldn't ride on an "if we survive" to get out of making good design choices.

There is also a more strategic reason not to treat present-day harms as someone else's problem: they are the only practice we get. The objection raised in class was that we only have one chance to get AGI alignment "right," and I think that makes

the case for working on current failures stronger, not weaker. If the final attempt cannot be repeated, then the skills it requires need to be developed somewhere recoverable. Today's systems are that place. The failures described above are small-scale versions of the same problem: a specification that conflicts with itself, a default user who does not match the real one, and no one having decided whose needs take priority when they collide. Working out how to elicit values from a diverse population, how to resolve disagreement among them, and how to audit whether a specification does what its authors intended are all things we can test now, while a mistake means a patch and not a catastrophe. A system that cannot be aligned to the needs of a teenager in distress is not one I would trust to be aligned to humanity's survival, and a field that has never practiced specifying values under real disagreement will be doing it for the first time when the stakes are highest. I do not want to overclaim here. Techniques that work on today's models may not scale to far more capable ones, and success on present harms is not proof of success later. But treating current societal harms as a trial run is still better than treating them as a distraction, because avoiding doom is itself a value specification problem, and no one has yet shown that we know how to do those well.

Beyond arguing that this work belongs in the alignment field, I want to suggest an interdisciplinary way of incorporating inclusivity into alignment research that is different from what's been suggested by these papers. There is value in participatory design, a study technique that is used heavily in the HCI space, when it comes to building a constitution for AI. Constitutional AI's biggest problem, according to Conitzer et al., is the question of who is designing the constitution.

Anthropic has already tried a version of this. They polled the public when they partnered with the Collective Intelligence Project in 2023, and surveyed 1000 American adults from diverse gender, race, economic, and geographic backgrounds to design a public constitution [5]. They asked the participants to vote on rules for their AI chatbots, either by voting on pre-existing rules, or by introducing new rules that other people could vote on. Following this, they took the rules and turned them into a public constitution that they did a Constitutional AI training pipeline on for a small model. This model was then compared to a Claude Instant model. This project had some interesting results. First was that the core alignment and capability was preserved– that is, it performed comparably on common AI language and math tasks, and evaluators found this public experimental model to be just as helpful and harmless as Claude models. Second, there was a measurable reduction in stereotype bias, especially when it came to disability and physical appearances. Third, there was a fifty percent overlap between this constitution and

Anthropic's actual constitution at the time, though the public constitution placed greater importance on objectivity, impartiality and accessibility. Finally, the public model was comparable to the Claude Instant model in terms of political alignment when measured against OpinionQA benchmarks [5].

I would like to point out that this model was not rolled out due to scalability issues (scaling democratic input is hard), and the study only surveyed 1000 American adults (which definitely does not include children, and probably does not involve many people with cognitive disabilities) in a rule-deliberation framework. Rule-deliberation is different than participatory design as the former has participants vote on statements while the latter has participants co-designing with others in a structured and accessible session. However, the big takeaway I would like to highlight is that it does suggest that public consultation can affect AI behavior (potentially for the better) without degrading AI capability, and this can influence future frameworks where public deliberation can be brought into the alignment problem.

Participatory design workshops could address the question of who designs the constitution, as you can sample the constituents of a product and ask them how they would design a contract for that product's use and what they want to see. If you get a representative sample of the general population, that includes more vulnerable populations such as children, the elderly and the disabled, you can survey this group and ask them to design contracts for how AI should operate. If you design a constitution with input from an inclusive pool of people, it would potentially do a better job at representing these people than a company can. I think participatory design is worth exploring, as the people who are affected by AI and AGI can co-design what they would like to see in their AGI product. Even if it doesn't give us a concrete solution due to scalability issues, it can give us a better idea of what aligned AI might look like in the public eye.

Whether participatory design is the most ideal approach for alignment with society is definitely debatable - I just think it's an approach worth exploring. Regardless of the debate on what is the best way to align with humans, figuring out how to represent and formalize humanity's needs (and wants) is an open and interdisciplinary problem worth examining now, rather than examining too late.

I would like to leave this note with the reader: whoever answers the question "aligned to whom?" is contributing to the definition of alignment. Even if it complicates things, that answer should include the people who will live with the result.

## REFERENCES


[1] Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H. Holliday, Bob M. Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, Emanuel Tewolde, and William S. Zwicker. Social choice should guide AI alignment in dealing with diverse human feedback. In *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235, 2024.

[2] Andrew Smart, Shazeda Ahmed, Jackie Kay, Jimmy Tobin, Kris Shrishak, and Abeba Birhane. Toward a theory of value in AI alignment. arXiv preprint arXiv:2608.10327, 2026.

[3] Sara Wachter-Boettcher. *Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech*. W. W. Norton, 2017.

[4] *Raine v. OpenAI, Inc.*, No. CGC-25-628528 (Cal. Super. Ct., S.F. County, filed Aug. 26, 2025).

[5] D. Ganguli et al. Collective Constitutional AI: Aligning a language model with public input. Anthropic, 2023. https://www.anthropic.com/index/collective-constitutional-ai-aligning-a-language-model-with-public-input

[6] Mohammad (Matt) Namvarpour, Brandon Brofsky, Jessica Y Medina, Mamtaj Akter, and Afsaneh Razi. 2026. Understanding Teen Overreliance on AI Companion Chatbots Through Self-Reported Reddit Narratives. In Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26). Association for Computing Machinery, New York, NY, USA, Article 1646, 1–19.

[7] Yudkowsky, E., et al. (2016). Value. LessWrong. https://www.lesswrong.com/w/value https://doi.org/10.1145/3772318.3790597