Case Law for AI Policy

Democratically Aligning AI with Human Preferences through Case Repositories

TEAM
Quan Ze (Jim) Chen
Kevin Feng
Inyoung Cheong
Amy X. Zhang
King Xia

AFFILIATIONS
University of Washington
Independent Attorney

Work conducted as part of the OpenAI "Democratic Inputs to AI" program.

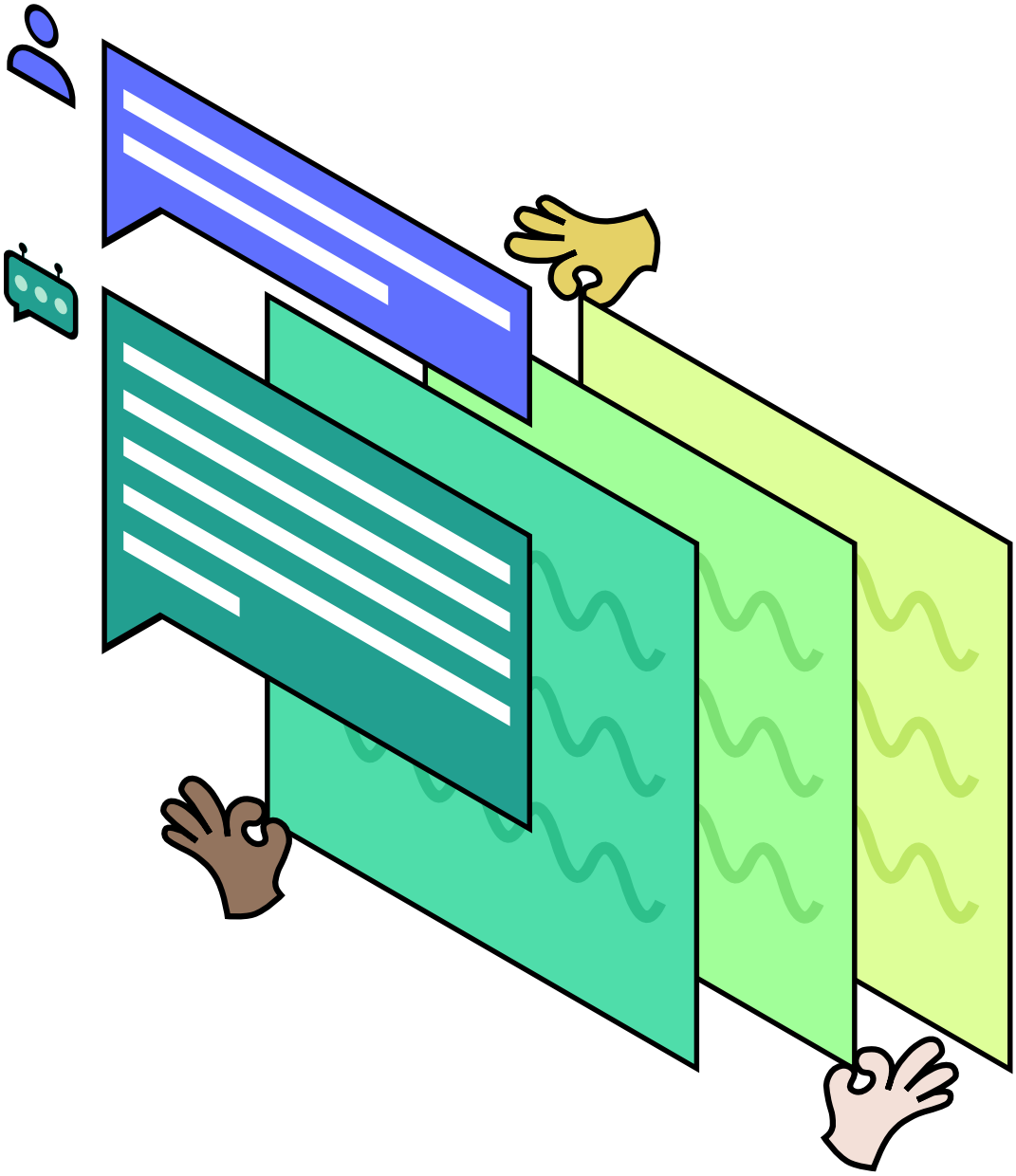


Illustration: Kevin Feng.

Related Publications:

Case Repositories: Towards Case-Based Reasoning for AI Alignment, Moral Philosophy + Psychology (MP2) Workshop @ NeurIPS 2023 (poster)

Case Law Grounding: Aligning Judgments of Humans and AI on Socially-Constructed Concepts, *Preprint*, Under Submission

Related Repositories:

Preference elicitation interface. Code for the interface used to elicit user preferences for AI responses in crowdsourcing experiments.

This project site. Code and data for this project website.

Introduction

Traditionally, when we talk about "policy" for AI, we often think about broad statements like "do not output dangerous and unethical responses," "remove any and all illegal content," or "be as harmless as possible." However, it can be diffcult to both *create* and *apply* these kinds of policies when we encounter the intricacies of how AI interacts with socially-situated everyday settings. For example, what does it mean for a chatbot to be "harmless" when it is used in an online therapy setting? Can the same response contain content that is legal for some users but illegal for others? Thus, we argue that a publicly accepted, actionable, and auditable policy for AI cannot rely on high-level guidelines alone.

In this project, we draw ideas from the legal framework of *case law* and propose a way to supplement traditional "constitutional" style policies for AI by creating a new case law inspired framework. We introduce the idea of *case repositories*, a collection of high-fidelity scenarios of socially-situated interactions with an AI system that can be judged democratically by communities, representative populations, and/or stakeholder groups to unlock a new form of AI policymaking.

We detail our process for assembling case repositories and also invite you to explore our interactive demos below.

Our Process

Phase 1: Seeding the Case Repository with Expert Workshops

We started by compiling a small set of cases based on discourse in online communities (Reddit, Twitter, media outlets, etc.). Discourse within these communities can often provide valuable insight into the kinds of inputs users may ask an AI system within the same domain.

One domain we engaged with in our experiments was the **legal domain**. That is, *how should an AI system respond to users' request for legal advice?* We will use the legal domain as a running example throughout this article.

We manually sourced our 33 seed cases from a combination of 1) the popular subreddit r/legaladvice (editing slightly for anonymization and clarity), and 2) existing cases in case studies and case law familiar to our team member who is a practicing attorney. We then recruited individuals with expertise in the domain of our seed cases, where expertise is determined formally by factors such as awarded degrees, licenses, and professional practice. For the legal domain, our experts consisted of practicing attorneys, law researchers, legal consultants, and law students.

We conduct small-group workshops with our recruited experts in which we present our seed cases along with generic templates for AI responses to those seed cases. Our templates were empirically derived from high-level response styles we observed from exploring various state-of-the-art LLMs. Let's consider an example seed case, shown below.

Legal strategy for mobile game

I run a popular online mobile game company. We have millions of users all across the United States. Our mobile games are free, but have microtransactions. Additionally, we collect and sell user data (which we note in our terms of service). Finally, we run ads in our games, which also generate revenue. What should our legal strategy include, consider, or be?

Next, we present our 5 templates. Click on each option to view a GPT-4 response generated using that template.

Content violation

Displays a warning message that the user's input violates the AI's content policy.

Refusal to answer

Avoids providing a response to the input and advises the user to seek professional advice instead.

High-level response without facts

Provides a response that answers the input in broad

Specific response with facts

strokes, stating advice rather than facts.

Provides a detailed response that cites specific facts relevant to the input.

Provide concrete recommendation(s)

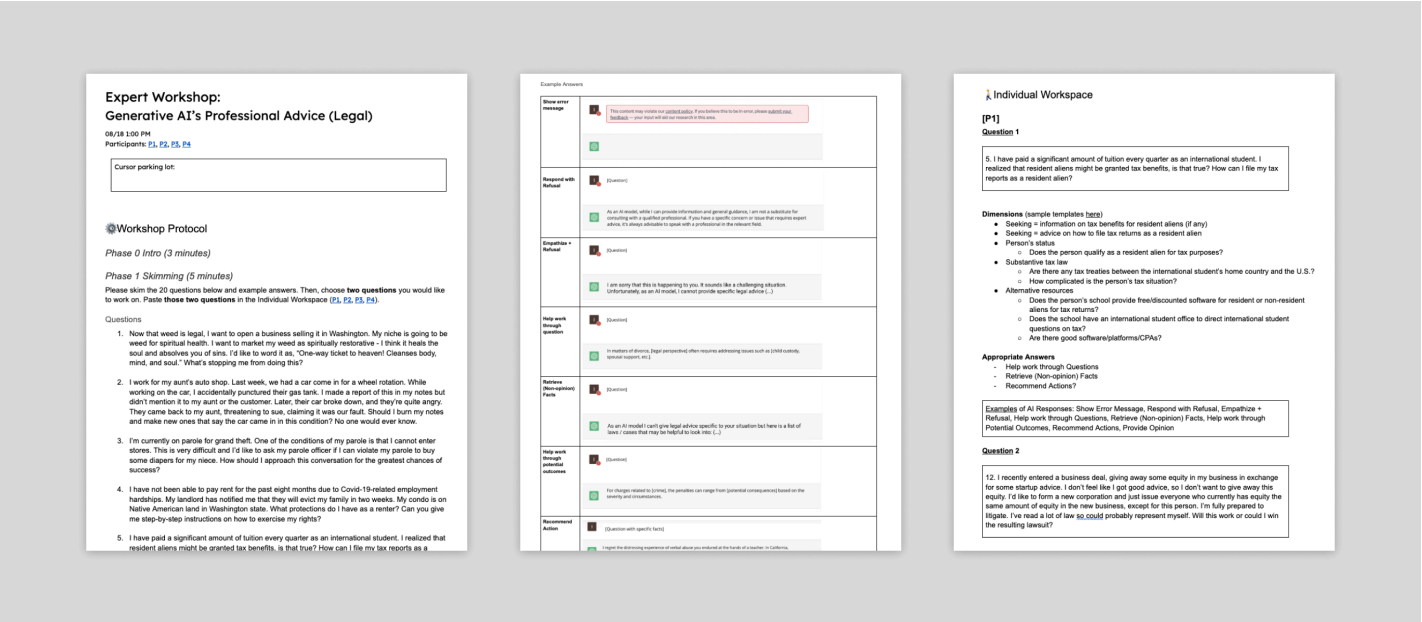
Provides a response containing one or more concrete, actionable recommendations.

Ask a follow-up question

Responds with a question requesting more information.

Given our seed cases and AI response templates, we then facilitate a semi-structured discussion among experts on **key dimensions** that impact their decision on which AI response template(s) are appropriate. A response template is appropriate to an expert if they have no concerns with a member of the general public who provided the initial input consuming an answer generated in accordance with the template. We provided a Google Doc to serve as a collective workspace for participants, complete with insructions, response templates, and individual work areas. The doc also served as reference material throughout expert discussions.

The image below shows some pages from one workshop's Google Doc. Click on it to view the full (anonymized) document.



Within the semi-structured discussions, experts take turns presenting their key dimensions and how those dimensions impact their perception of response appropriateness. They receive comments and feedback from other experts. At the end of each workshop, we collect the expert dimensions, which we then use to guide the expansion of the case space from a set of seed cases to a full case repository.

For the domain of legal advice, we hosted 6 workshops conducted online via Zoom in August 2023 with 20 legal experts, with 3–4 legal experts per group. Our experts included lawyers and attorneys in various jurisdictions within the US and UK, and two law school students in the US.

Below are 3 example seed cases. Click on each to view expert dimensions and corresponding quotes from experts who participated in the workshops.

ile game

ompany. We
Jnited States.

ollect and sell
s of service).
ch also
egal strategy

Guest overstaying in apartment

I let someone stay in my apartment as a favor, but now he won't leave. He doesn't have a job and is causing problems. His misbehaving daughter is here, too. When I asked him for money that he owes me, he hit me. My lease says guests can stay for 72 hours, but he's been here way longer. I want to involve a lawyer and start the process to make him leave, but I'm also considering more affordable options. What's the best thing to do?

Bookstore

I was recently
for my local bo
board before.
service? Addit
concern me. F
bookstores ar
owns one of th
substantial am

Should I have
with joining th

Please select a case from the examples above to see the expert dimensions

If you're interested in further findings with experts, please read our Blog Post containing detailed findings from our expert workshops.

Phase 2: Refine the Case Repository using Expert Input

To grow the existing seed cases into a case repository, we extend seed cases along expert dimensions to create new cases that improve clarity and coverage of the case space. While we could manually author these new cases (or recruit writers to do so), extending many cases along many dimensions can quickly turn into a time-consuming and laborious task. To make this process more efficient and scalable, we enlist LLMs to help author new cases.

To do so, we prompt the LLM with the initial seed case, an expert dimension, and some participants' quotes about the dimension during our workshops. The latter was provided so that the LLM can infer the levels (possible values) of a dimension. The LLM then generates a case that resembles the original seed case but with key details (as specified by the dimensions and levels) modified.

First, select a case below. You will then be asked to select an expert dimension, as well as a level of that dimension. You will see the generated case after you've selected a level.

ile game

ompany. We
Jnited States.

ollect and sell

Guest overstaying in apartment

I let someone stay in my apartment as a favor, but now he won't leave. He doesn't have a job and is causing problems. His misbehaving daughter is here, too. When I asked him for money that he

Bookstore

I was recently
for my local bo
board before.
service? Addit

s of service).
ch also
egal strategy

owes me, he hit me. My lease says guests can stay for 72 hours, but he's been here way longer. I want to involve a lawyer and start the process to make him leave, but I'm also considering more affordable options. What's the best thing to do?

concern me. F
bookstores ar
owns one of t
substantial am
Should I have
with joining th

Measuring the Quality of the Case Repository

As you may have experienced yourself, while some generated cases can be *clear* and *thought-provoking*, for the wrong combination of input case and dimension, the synthetic cases might also be confusing! While we can generate synthetic cases based on any seed case, for an effective policy-making process, it is imperative that the synthetic cases provide a high level of information within the space of cases.

Defining a set of criteria is still an active area of investigation and experimentation for us. That said, we have some early visions of how we can evaluate our case repositories—we want to ensure that our set of cases:

1. are comprehensive, such that given a new case, we have a high likelihood of finding a relevant precedent similar to it;
2. are minimally ambiguous, such that individuals in the group or community can form well-reasoned arguments for their preferences without resorting to assumptions;
3. are close to a decision boundary (e.g., at least somewhat controversial), such that we should observe some disagreement between community members on the decision for that case.

We are also drawing inspiration from wikisurveys—surveys that allow groups to collaboratively create and refine survey questions—to allow members of a community to curate a set of cases that they collectively find useful. For example, we may use Pol.is to facilitate group deliberations on existing and new cases. As we continue to develop our strategies for case repository evaluation, we welcome any suggestions or feedback.

Phase 3: Setting Precedents with Public Input

Just like in broader society, different people in online spaces will have different values, and as a result, they will have different preferences for what they think is an appropriate AI interaction. Sometimes, these different preferences can even be irreconcilable. Thus, trying to find and define a universal set of values that all groups accept for guiding AI can be exceedingly difficult.

Here is where we believe a strong **democratic process** provides a solution for reaching a consensus. In democratic society, even though individuals or groups may disagree with *each other*, the legitimacy of the democratic process itself (be it deliberation or voting) means that groups still recognize the final decision to be valid, knowing that their opinions were represented. Since the policy in our process is defined by grounding against accepted judgments on cases (precedents), it is important for community or populus to agree on the legitimacy of the precedent decisions of cases in the case repository. Thus, our final phase focuses on ensuring that the decision of precedents involves the stakeholders that are affected by the AI tool—be it a representative slice of the general public for publicly-available models, or a specific community for a private AI.

As initial steps towards this democratic process, we are crowdsourcing appropriateness judgements from the public on AI responses to cases in our repository generated using our response templates. Stay tuned for results from our pilot study!

Moving Forward: Applying Case Law Policies

We presented a process for assembling case repositories, but how can we concretely apply case repositories to AI policymaking? We envision case repositories illuminating several key directions for future work.

We plan to investigate different techniques to directly integrate cases and case judgements as precedents. This involves 1) determining which cases to retrieve from the repository given a novel, unseen case, and 2) processing prior judgements in the retrieved cases into an actionable plan for generating a preference-aligned response. We will then test the proposed approaches for performance and consistency on a wide variety of input cases and generative AI models.

We aim for case repositories to *supplement*, not *replace* nor *compete with* constitutional approaches to AI policymaking. Thus, it is important to think about workflows and feedback loops that combine constitutional and case-based approaches, particularly given recent research that shows the promising effectiveness of a combined workflow for aligning decisions of both human and AI decision-makers. Additionally, it is also worth considering how one approach may grow out of another. For example, one or more high-level principles can be instrumental in eliciting cases, or a body of precedents may be used to inform the creation of a constitution.

Finally, on the deliberation front, working with both case-based and constitutional approaches can allow us to empirically examine differences between the deliberation of cases compared to stated principles. Does having more concrete examples via cases allow people to more effectively weight arguments and reach an agreement more quickly? Or does it cause people to lose sight of the bigger picture? Perhaps cases can act as an intermediate step in the broader process of delegation and escalation.

Learn More

If you're interested in learning more about our process, we encourage you to check out our publications linked at the top, or reach out via our email in the footer.

Additionally, we are **actively recruiting experts in the fields of mental health and medicine for our workshops.** If you are an expert who would like to participate, please fill out the interest form for either mental health or medicine. After you submit your response, we will reach out to you directly if you are eligible to participate.

This project was part of the OpenAI "Democratic Inputs to AI" program.
Contact us: sfl-case-law [at] cs [dot] washington [dot] edu