PolicyCraft: Supporting Collaborative and Participatory Policy Design through Case-Grounded Deliberation

Tzu-Sheng Kuo
Carnegie Mellon University
Pittsburgh, PA, USA
tzushenk@cs.cmu.edu

Quan Ze Chen
University of Washington
Seattle, WA, USA
cqz@cs.washington.edu

Amy X. Zhang
University of Washington
Seattle, WA, USA
axz@cs.uw.edu

Jane Hsieh
Carnegie Mellon University
Pittsburgh, PA, USA
jhsieh2@cs.cmu.edu

Haiyi Zhu*
Carnegie Mellon University
Pittsburgh, PA, USA
haiyiz@cs.cmu.edu

Kenneth Holstein*
Carnegie Mellon University
Pittsburgh, PA, USA
kjholste@cs.cmu.edu

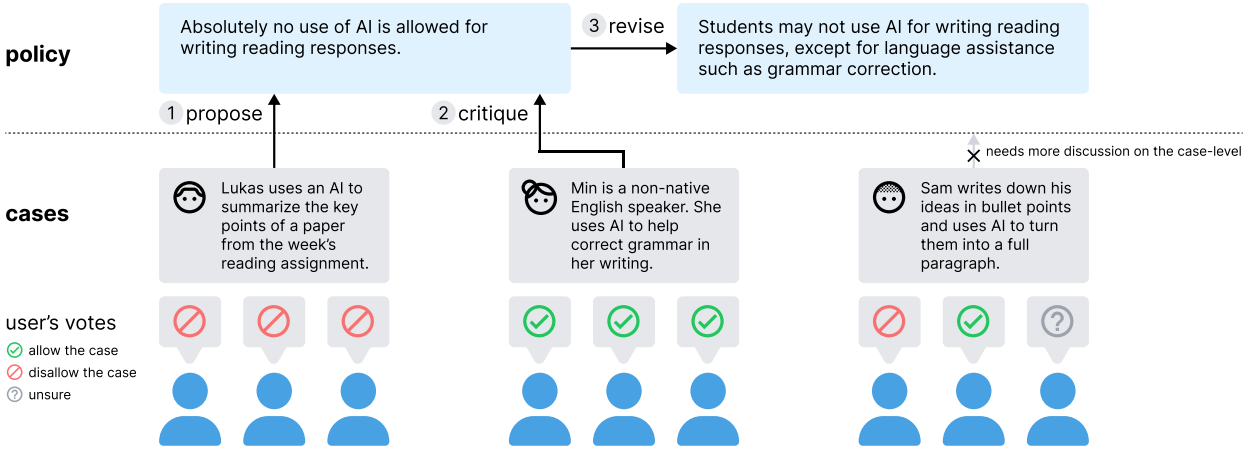


Figure 1: PolicyCraft is a system that supports users in collaboratively proposing, critiquing, and revising policies through discussion and voting on concrete cases. Throughout the iterative policy design process, PolicyCraft helps users understand whether they disagree about the wording of a policy or whether they have an underlying disagreement about how specific cases should be handled. The system then supports users in discussing and addressing disagreements accordingly.

Abstract

Community and organizational policies are typically designed in a top-down, centralized fashion, with limited input from impacted stakeholders. This can result in policies that are misaligned with community needs or perceived as illegitimate. How can we support more collaborative, participatory approaches to policy design? In this paper, we present PolicyCraft, a system that structures collaborative policy design through *case-grounded deliberation*. Building on past research that highlights the value of concrete cases in establishing common ground, PolicyCraft supports users in collaboratively proposing, critiquing, and revising policies through discussion and voting on cases. A field study across two university courses showed that students using PolicyCraft reached greater consensus and developed better-supported course policies, compared with those using a baseline system that did not scaffold their use of concrete cases. Reflecting on our findings, we discuss opportunities for future HCI systems to help groups more effectively bridge between abstract policies and concrete cases.

CCS Concepts

• **Human-centered computing** → **Collaborative and social computing systems and tools**; • **Social and professional topics** → **Computing / technology policy**.

Keywords

policy, deliberation, case-based reasoning, participatory design, AI

ACM Reference Format:
Tzu-Sheng Kuo, Quan Ze Chen, Amy X. Zhang, Jane Hsieh, Haiyi Zhu, and Kenneth Holstein. 2025. PolicyCraft: Supporting Collaborative and Participatory Policy Design through Case-Grounded Deliberation. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 24 pages. https://doi.org/10.1145/3706598.3713865

*Co-senior authors contributed equally.

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713865

1 Introduction

Communities and organizations of all types, whether small or large, online or offline, often rely on *regulatory policies*[1] to support community governance [59]. For example, online communities like Reddit have content moderation rules that govern what content is allowed in each subreddit [14, 29]. Small groups like university classes have course policies that guide how students may or may not use generative AI in their coursework [85]. Local governments have zoning regulations that determine what kinds of buildings can or cannot be developed in specific geographic areas [44], such as restricting industrial development in residential neighborhoods. Such regulatory policies are essential for guiding the behavior of community members, upholding their shared values, and supporting their collective goals [68].

Today, policy development processes are typically top-down and centralized, lacking input from the communities and stakeholders who they will impact [39, 80, 101]. This can lead to the development of policies that are poorly fit to community needs [97]. Furthermore, community members may perceive these policies as illegitimate [101], leading them to reject the policies [9], organize strikes [62], or even leave the community [28]. For example, volunteer moderators on Stack Overflow organized a strike to protest a new policy about moderation of AI-generated content, which the platform enforced without first communicating with the community [62].

While more participatory, bottom-up approaches to policy design hold potential to support greater alignment with community needs and values, it can be challenging to realize this goal in practice. Even when communities share overall norms and values [14, 29], individual community members may hold differing perspectives [22, 52, 92]. Given that policy proposals are often high-level and abstract, it can be challenging for community members to identify the root sources of their disagreements and to collaboratively refine policies to resolve these disagreements [17].

Past research has shown that the use of concrete cases or scenarios is critical for establishing common ground in discussions about policy design [16–18, 27, 45, 52]. However, connections between concrete cases and abstract policies are often made in an ad-hoc and inconsistent manner during policy design conversations [13, 23, 48, 69, 72]. For example, while people are often inspired to propose policies based on specific, concrete scenarios (whether real or imagined), they do not always discuss how proposed policies might have unintended consequences in *other* scenarios [23, 48, 55, 94]. Similarly, while people often iterate on the wording of policies to better address imagined scenarios or edge cases, the cases themselves are not always explicitly communicated as objects for discussion [23, 34, 35, 69, 94].

In this paper, we present PolicyCraft, a system designed to structure collaborative policy design through *case-grounded deliberation*. As illustrated in Figure 1, PolicyCraft supports users in collaboratively proposing, critiquing, and revising policies through discussion and voting on concrete cases. In doing so, PolicyCraft is designed to help users pinpoint where their disagreements come from—whether they are about the wording of a policy or a more fundamental disagreement about how specific cases should be handled—and then discuss and address these disagreements accordingly. Using PolicyCraft, users share and discuss their perspectives on whether specific concrete cases should be allowed or disallowed, regardless of what current policies say. Building upon consensus at the case-level, users *propose* new policies or *critique* and *refine* current policies in order to better reflect their collective viewpoints.

We evaluated PolicyCraft through a field study across two university classes, where students used PolicyCraft to collaboratively design course policies regarding which generative AI use cases should be allowed in their classes. We find that policies developed using PolicyCraft received stronger support and achieved greater consensus among the students who developed them, compared with a baseline system that did not scaffold students in using concrete cases to support policy design. Students using PolicyCraft also found it easier to understand each other's perspectives, and were more motivated to iterate on policies based on cases shared and discussed by others. The policies developed by students in our study were refined by course instructors and students and adopted as official course policies.

Overall, this work contributes PolicyCraft, a system that supports collaborative and participatory policy design through the systematic use of cases. We also show that policies created with PolicyCraft receive stronger support and consensus from those involved in the policy development process. Reflecting on our findings, we discuss opportunities for future HCI research and design to support groups in more effectively bridging between abstract policies and concrete cases during collaborative policy design.

2 Related Work

Policy design has been a key area of focus in HCI research [37, 54, 83, 96]. Following prior HCI scholarship, in this paper we understand policies as principles, guidelines, and written rules to guide behavior in communities and organizations [37, 96]. From investigating how policies are designed and implemented in practice [10, 12, 29, 90] to creating tools that help develop policies [51, 99], HCI researchers view policy design as a critical topic of inquiry because policies directly impact how technology shapes communities and society [37, 96]. In this section, we first discuss the importance of collaborative and participatory approaches to policy design. We then review existing tools that aim to support such approaches. Finally, we introduce the concept of case-based reasoning [1], which inspires the design of PolicyCraft.

2.1 Collaborative and Participatory Approaches to Policy Design

Policies often face a crisis of legitimacy because they don't meet the needs and values of impacted communities. For example, online social platforms are often criticized as arbitrary and censoring free speech because their centralized trust and safety team enforces content moderation policies without community input [39, 101]. Elected governments that impose top-down policies solely based on their own agendas also risk losing public support and future elections [6, 73]. To address the legitimacy crisis, researchers suggest adopting more collaborative and participatory approaches to policy design [19, 39, 50, 68, 73, 86, 90, 93, 101]. As Ostrom stated

[1]Regulatory policies focus on regulating the behavior and practices of individuals within communities and organizations, while other policies like constituent, distributive, and redistributive focus on how resources, services, or responsibilities are shared or organized [59].

in her influential principles on governing the commons [68]: *"individuals affected by the operational rules should be able to participate in modifying those rules."* By incorporating the local knowledge of impacted individuals and communities, collaborative and participatory approaches ensure that rules and policies are well-suited to local circumstances [90]. These approaches also help stakeholders view the policies as legitimate, even if they disagree, because of the inclusive and deliberative way in which they were settled [100]. Policies developed collaboratively by impacted communities are essential for maintaining stability, building trust and legitimacy, and reflecting collective goals and values [68].

However, engaging the community in collaborative and participatory policy design poses several open challenges. First, even within a community with shared norms and values, individual members may still have differing and often conflicting perspectives [73]. For example, on Wikipedia, even with clear policies on vandalism, individual Wikipedians can still interpret and apply the policy differently when moderating article edits [31, 52]. These different interpretations exist because policies are often high-level and abstract, making it difficult for people to pinpoint the source of their disagreements without referring to concrete examples [17]. Besides accounting for different community perspectives, another challenge arises when community members collaborate on policy iterations. Without a structured, coordinated process, it can be challenging for community members to achieve meaningful collective action, even if they are eager to contribute [77]. Finally, individual community members often have different preferences, availability, and capacity, and thus require varying levels of support to participate effectively [9, 24, 53, 87]. These challenges emphasize the need for research into tools and processes to support collaborative and participatory policy design, which we discuss in the next subsection [37, 96].

2.2 Tools for Collaborative and Participatory Policy Design

Existing tools for collaborative and participatory policy design range from those that mainly gather community perspectives to *inform* policy design to those that directly engage participants in drafting policies [5]. In the first category, researchers have developed various tools to gather community and public perspectives. For example, Polis is an online platform where participants submit short statements expressing their views on a topic and vote to agree or disagree with statements from others [81]. It is widely used in policy design worldwide, such as by the Taiwanese government to craft UberX regulations based on input from citizens, taxi drivers, Uber Inc., and other stakeholders [33]. ConsiderIt is another prominent platform where people submit the pros and cons of a policy and compare their views with those of others who agree or disagree with the policy [51]. The City of Seattle has used ConsiderIt to gather public input on policy proposals, such as the plan to increase affordable housing in certain residential areas. Several other tools use visualizations [42, 58, 61, 97] or chatbots [43, 61, 78] to identify disagreements for consensus building.

In the second category are tools that allow individuals and communities to contribute to the drafting of actual policies. For example, drawing inspiration from microtasks in crowdsourcing [56], researchers developed CommunityCrit, a system where users can propose policies and comment on others' policy proposals through micro-activities [60]. Using CommunityCrit, the researchers worked with a local planning group in San Diego to redesign an intersection based on crowdsourced proposals [38]. Other works have crowdsourced policy demands against entities responsible for privacy or labor rights violations [75, 95]. For example, building on the find-fix-verify crowd programming pattern [8], Wu et al. designed questionnaires that ask independent crowd workers to find privacy concerns, propose potential fixes, and verify and rank the proposed fixes [95]. With the rise of generative AI, some tools like Remesh use GPT-4 to automatically synthesize crowdsourced public views into initial policies and then iteratively refine them through expert and public feedback [47]. Other tools like PolicyKit and Pika provide infrastructures that allow online community members to set up their own process for policy design [91, 99]. Overall, these tools enable individuals to share and gather feedback on policy ideas, or to contribute inputs for others to synthesize into policies. However, they are not focused on supporting communities in directly collaborating to iteratively and deliberatively craft policy proposals.

PolicyCraft bridges between these two categories of tools by supporting communities in both *sharing perspectives* and collaboratively *authoring policies*, in an integrated, iterative workflow. Unlike Polis, ConsiderIt, and other tools in the first category that mainly focus on gathering participants' perspectives to inform policymakers, PolicyCraft further empowers participants to draft actual policies based on the cases they vote on and discuss. This elevates participants from a consultative role to a more collaborative role in policy design [5]. Meanwhile, PolicyCraft differs from related work in the second category by supporting groups in directly collaborating and deliberating on policy drafts. To scaffold participants in doing so more systematically, PolicyCraft promotes *case-grounded deliberation*, structuring users' discussions and contributions around the dual abstractions of *cases* and *policies*. This structuring aims to support both individuals and groups in effectively transitioning between case-level and policy-level discussions during policy design—including by helping them tease apart whether they disagree at the case-level versus the policy-level, and by helping ground policy iterations in consensus on concrete cases.

2.3 Cases as a Medium for Policy Design and Deliberation

Cases have been used as a medium for reasoning and deliberation across many fields [1]. Following prior HCI scholarship and literature in case-based reasoning, in this paper we understand cases as concrete scenarios within specific problem situations that make abstract concepts tangible for reasoning, communication, and negotiation [1, 16, 22, 27, 35]. For example, cases involving ethical dilemmas, such as the trolley problem [89], are frequently used to guide moral deliberation and explore ethical principles [7, 63, 66]. Legal theories also often rely upon case-by-case judgments and deliberations to guide decision-making, rather than directly applying top-down rules [11, 18, 30]. Inspired by jury trials in the legal domain, HCI researchers have proposed case-based discussion tools such as digital juries to adjudicate content moderation cases [25, 49]. While these tools do not surface cases to the policy level for

policymaking, the concrete cases offer a common ground that enables shared understanding and facilitates meaningful deliberation around what decisions and behaviors are desirable [16, 27].

Research has shown that the use of concrete cases plays a major role in supporting effective policy design [23, 55, 94]. For example, Wikipedians deliberate on and iteratively refine their definition of vandalism based on specific article edits they come across while moderating Wikipedia articles [31, 52]. Other work also uses concrete vignettes and scenarios to support stakeholder reflection and deliberation around the ethical, legal, and policy implications of emerging technologies (e.g., [21, 41]). These concrete cases help people identify the source of their disagreements, whether it's due to ambiguous policies or genuine differences in their perspectives about how specific cases should be handled [17]. When making collective decisions, deliberation grounded on concrete cases further provides procedural legitimacy and helps build consensus, even in the face of disagreements [25].

However, connections between concrete cases and abstract policies are currently made in an ad-hoc and inconsistent manner during policy design [13, 23, 48, 69, 72]. For example, people often propose policies based on specific cases and scenarios, but they do not always discuss how these policies might create unintended effects in *other* scenarios [23, 55, 94]. This can cause issues, as in online content moderation, where people usually agree with the general rules proposed based on past incidents but often feel frustrated when those rules are applied to situations they had not previously considered [48]. Meanwhile, although people with situated knowledge of the policy context often provide policy suggestions based on cases grounded in their lived experiences, the cases themselves—which provide important rationale for the policy suggestions—are not often explicitly shared in policy discussions [23, 34, 35, 69, 94]. Researchers have called for a more systematic approach to constructing and using concrete cases to support policy deliberation [66]. For example, researchers have introduced the concept of *research through litigation*, which involves carefully selecting cases to surface serious concerns and drive policy change [45]. PolicyCraft aims to better support this process in the context of collaborative and participatory policy design, by enabling users to create and iterate on policies through the systematic use of cases.

3 Design Goals

We established the design goals for PolicyCraft based on our review of prior work in Section 2. To ensure that these design goals aligned with real-world needs for community participation in policy design, we then validated them through one-hour, semi-structured interviews with six community organizers who had previously engaged in community policy design. These included two senior moderators responsible for shaping content moderation policies on Wikipedia and Stack Overflow, and four course- and university-level policy designers at a major US university. The discussion topics used to guide these semi-structured interviews can be found in Appendix A. Overall, we distilled the following four design goals for systems that aim to support collaborative policy design.

D1. The system should encourage users to develop policies based on concrete cases. Given evidence that discussing concrete cases can help people establish common ground during policy design conversations [23, 52, 55, 94], systems should support users in making systematic use of cases. Systems could draw inspiration from prior work, such as the idea of research through litigation [45, 96], to support users in using cases to surface potential flaws in current policies. These cases can provide users with shared context for policy development.

D2. The system should help users identify and address underlying sources of disagreements. People may disagree with each other during policy design for various reasons [73]. For example, they may substantively disagree about how different cases should be handled, or they may simply have differing interpretations of a policy's wording [17, 52]. Systems should help users distinguish whether they disagree at the policy-level, case-level, or both. Systems should further assist users in discussing and addressing disagreements at the appropriate level.

D3. The system should support users in easily building upon and referencing each other's work. While community-driven approaches to policy development ensure the policies are better aligned with local community perspectives [68], it can be challenging for community members to achieve meaningful collective action without a coordinated process [77]. Systems should make it easy for users to build upon each other's contributions, leveraging their complementary knowledge, experiences, and backgrounds [73, 75]. Systems should also enable users to track how others build on their contributions and to collaborate on planning future actions [77].

D4. The system should accommodate different levels of engagement in policy design. People have different amounts of time, preferences, and capacities for participating in the policy development process [9, 24]. Systems should offer different ways for users to participate, including light-weight options like voting or more involved ones like creating a new policy [60]. For more involved tasks like editing or creating policies, systems should offer additional scaffolding and assistance to reduce barriers to participation.

4 PolicyCraft

Based on these design goals, we developed PolicyCraft, a web-based system that supports collaborative and participatory policy design through case-grounded deliberation. PolicyCraft is designed to support policy design across a broad range of community contexts. PolicyCraft focuses on supporting the design of *regulatory* policies [59]: policies that guide what is allowed or disallowed within a community context. For example, in online communities like subreddits [39], we envision that PolicyCraft may assist a community in designing its content moderation rules (policies) based on community members' posts (cases). In local communities such as vTaiwan [33], PolicyCraft may help refine regulations for emerging technologies (policies), such as the use of UberX in specific local contexts (cases).

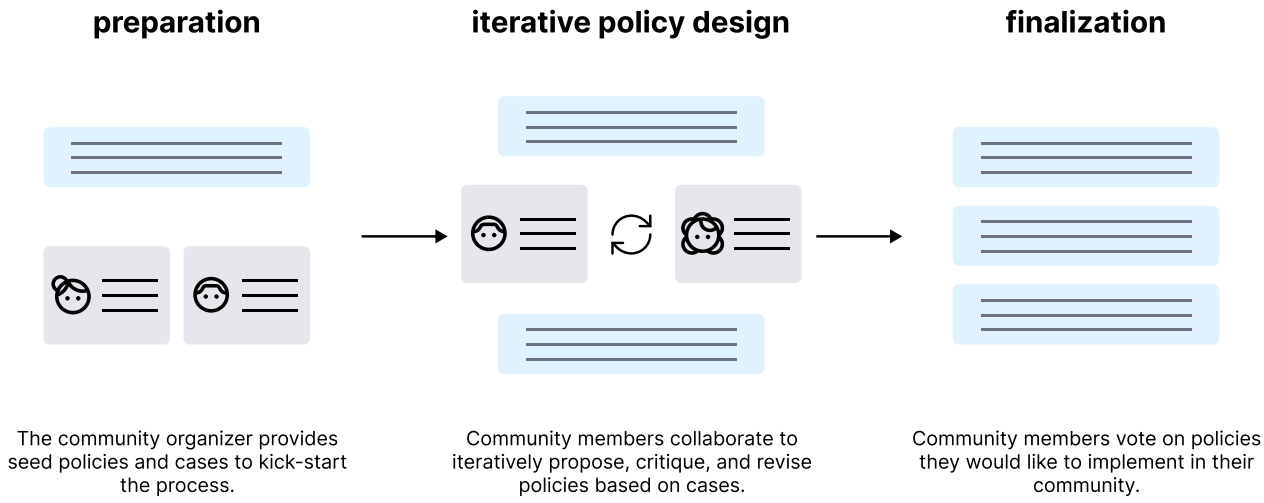


Figure 2: PolicyCraft's overall process.

4.1 System Overview

PolicyCraft supports policy design through the overall process illustrated in Figure 2. Community organizers start by entering a few initial seed policies and cases into the system. These seed policies/cases can serve to minimize the cold-start problem and establish norms around aspects such as length, formatting, and level of abstraction [15, 40]. Community members then collaborate to iteratively critique and revise the initial policies or propose new policies based on concrete cases. During this stage, community members can share their perspectives on whether specific concrete cases *should* be allowed or disallowed, regardless of what current policies say, through case-level voting and discussion. In the final stage, community members consider which of the policies they believe should be implemented in their community. They express their perspectives by anonymously upvoting or downvoting policies and then optionally providing publicly visible reasons for their votes. During the voting process, policies cannot be edited. This overall process is flexible and allows for customization based on community-specific needs and goals.

In the following subsections, we describe PolicyCraft's core functionality: supporting users in collaboratively *critiquing*, *revising*, and *proposing* policies through cases. We illustrate this functionality through a running example, describing how PolicyCraft would be used to support collaborative policy design in a university classroom setting. In this setting, students and instructors collaboratively develop course policies regarding which ways of using generative AI should be allowed in the course. The instructor serves as the community organizer and the students are community members. We also share additional features that help lower participation barriers and facilitate collaboration. Throughout the section, we connect specific features of PolicyCraft's design to the design goals outlined in the previous section, denoted as D1 to D4. Finally, we conclude with implementation details.

4.2 Critiquing Policies Through Cases

4.2.1 ***Users can critique a policy by creating cases that highlight ambiguities or potential flaws.*** After logging into PolicyCraft through their browser, users can start by going to the **policy repository**, where they can see an overview of all current policies. If users find a particular policy they think can be improved, they can click on it to see an expanded view on its **policy page**, as shown in Figure 3. As a concrete example, consider a user reading a course policy about whether AI use is permitted for coding assignments. The current policy says: *"Students may freely use AI for coding assignments with appropriate attribution."* Users can critique a policy by creating a *case*: a description of a concrete usage scenario that highlights ambiguities or potential flaws of the current policy (D1). For example, as illustrated in Figure 4 to highlight a potential flaw, a user may create a case that they believe is *currently permitted* under the policy, but which they believe *an ideal policy should disallow*.

When the user creates and adds a case to the policy, they **label** whether they think the current draft of the policy would *allow* or *disallow* the given case. Alternatively, if they think the policy is ambiguous regarding how the case should be treated, they can label it as *ambiguous*. Meanwhile, they also cast their personal **vote** on the case to indicate whether they think it ideally should be *allowed* or *disallowed*, or whether they are currently *unsure*. Finally, users provide a brief **reason** for their vote.[2] Note that votes and reasons are meant to reflect a user's personal opinion on whether a case should be allowed, regardless of what current policies say.

Once a user has added a case, it becomes visible in the **related cases** section of the policy page. Other users can then (1) edit the case's label if they have a different interpretation; (2) remove the case if they believe it is not relevant to the given policy; or (3) add additional related cases to the policy. The added case also

[2]We require users to provide a reason for their vote whenever they create a new case and vote to allow or disallow it. This design builds upon prior research showing that disagreement contributes to group ideation most when people elaborate and justify their stances [2]. Inspired by the design of ConsiderIt [51], if a user votes that they are "unsure", they can still explain their reasoning by providing separate allow and disallow reasons representing pros and cons they see.

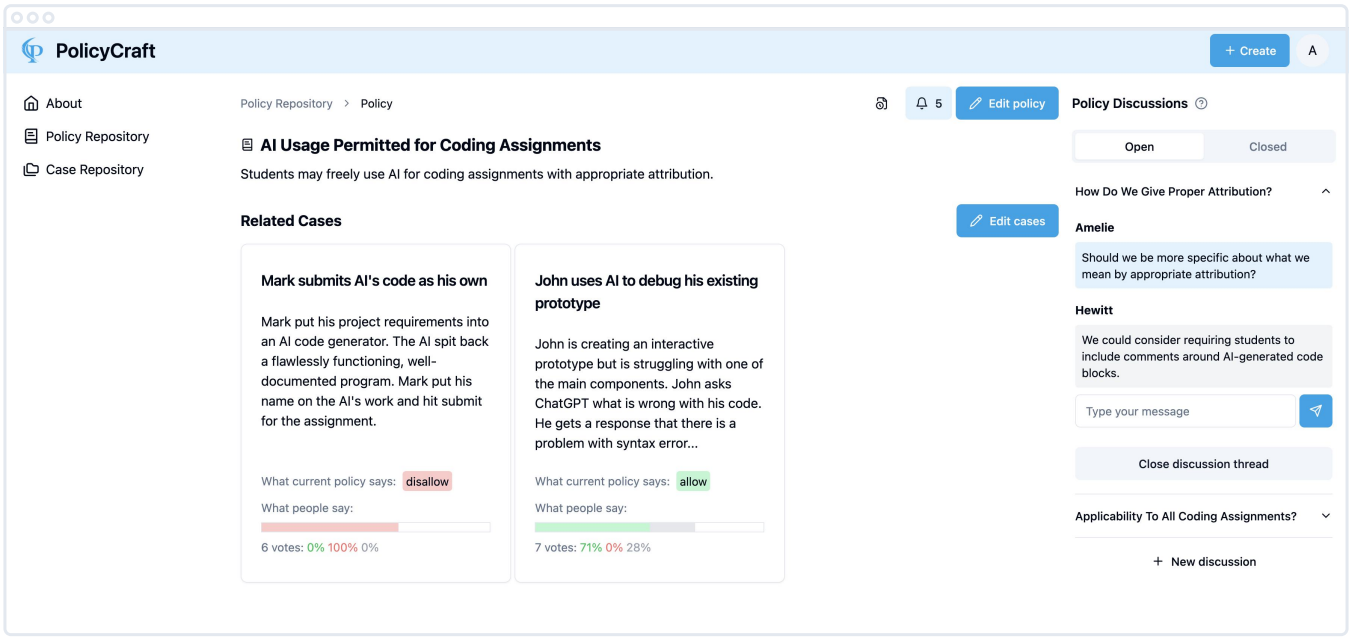


Figure 3: PolicyCraft's "policy page" for a given policy. On the left side is the navigation bar, where users can click to visit the Policy Repository to see all current policies, and the Case Repository to see all cases. Users can also visit the About page to view information about the current policy development campaign and to discuss general topics, not specific to a particular policy or case. In the center of the screen, the title and description of the selected policy is shown, along with its related cases. The visualization of users' votes on the cards is inspired by the design of Polis [81]. The bar represents how many people have voted out of the total number of users, while the percentages reflect the distribution of votes—whether to allow, disallow, or indicate unsure—among those who have already voted. Users may add new related cases by clicking the "Edit cases" button, which brings them to the page shown in Figure 4. They can also revise the policy by clicking the "Edit policy" button, which brings them to the page shown in Figure 6. Users can also start a discussion thread about this policy and close it once the topic has been resolved. Finally, users can propose new policies by clicking the "Create" button in the top-right corner.

becomes available in the **case repository**, where users can browse all created cases.[3] Anyone can read these cases and vote on whether they believe each case should ideally be allowed or not.[4] These votes and reasons help users understand what others think about specific cases (D2, D3).

4.2.2 ***PolicyCraft highlights misalignments between case labels and users' votes.*** Consider another user who visits the same policy page later and reads the case shown in Figure 4. Like the previous user, many others have also voted to disallow the case, although the current policy would allow it. As shown in Figure 5, whenever the label linking a case and a policy is misaligned with the majority vote on a case, a yellow alert message automatically appears to highlight it as a potential issue. In such cases of misalignment, the yellow alert message will suggest that *"The policy may need editing to better align with the majority vote on this case."* If most people vote to either allow or disallow the case, but the label indicates that the current policy is "ambiguous" with regard to whether that case should be allowed/disallowed, the yellow alert message will suggest that *"The policy may need editing to clarify whether this case is allowed or not."* These alert messages are intended to help people focus their policy iteration and discussion around cases where there is a misalignment between what the policy says about a case and what people believe (D1). If most people vote that they are "unsure" about whether a given case should be allowed, an alert message will not appear, regardless of the label, because this indicates that more discussion is needed at the case-level before using it to iterate on policy (D2).

4.3 Revising Policies Through Cases

4.3.1 ***Users can revise a policy to better align it with people's votes on cases.*** Upon noticing a misalignment between what a policy says and what people believe at the case-level (yellow alert messages), a user may decide to revise the policy by clicking the **edit policy** button on the policy page. This takes them to the policy editing page, as shown in Figure 6, where they can edit the policy in two steps. First, users can edit the policy's title and description as needed to better align it with people's votes on cases (D1). Then,

[3]The same case can be added to multiple policies as a related case. The label of a case represents a link between that case and a specific policy, so labels may vary across different policies. By contrast, users' votes are attached only to the case, independent of current policies.

[4]Users only see other people's votes after they cast their own. This design helps encourage more participation and prevent lurkers [67]. Cases cannot be edited, so people's votes stay relevant.

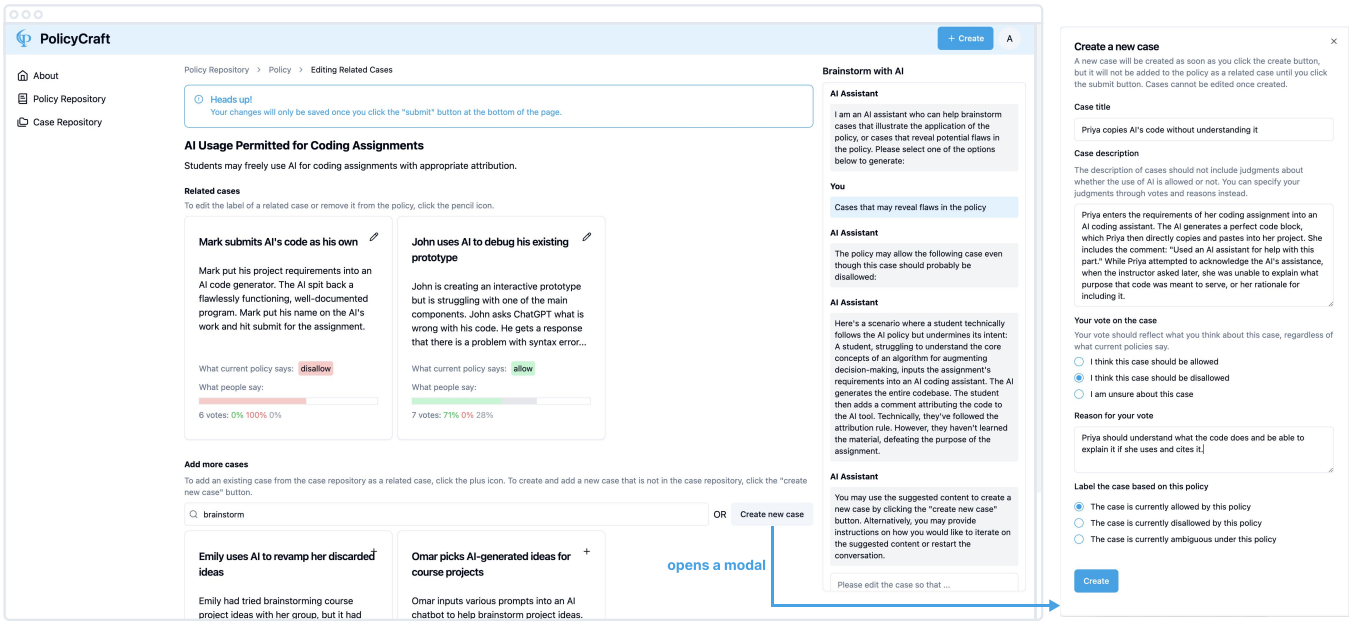
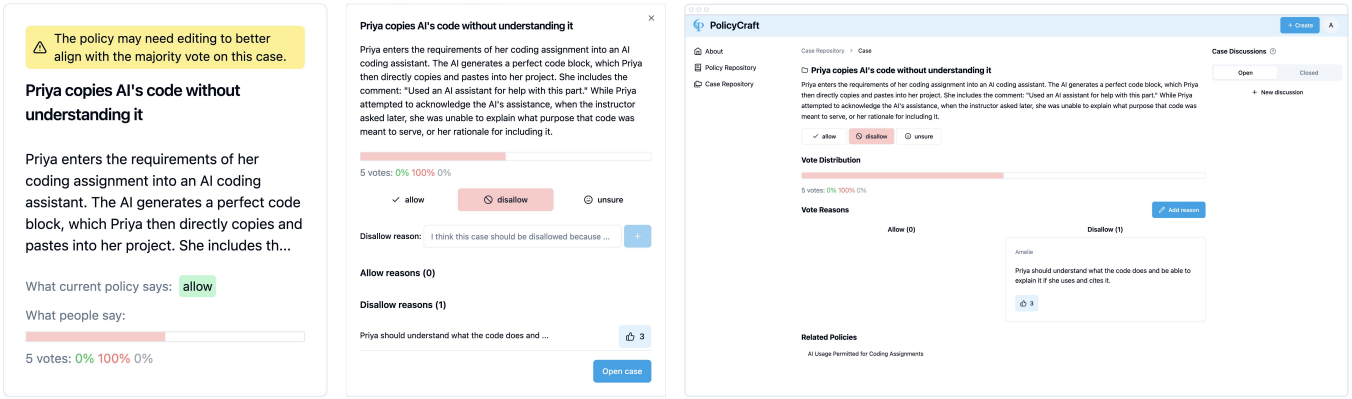


Figure 4: The workflow for editing a policy's "related cases" in PolicyCraft. The upper half of the screen shows the cases currently associated with a policy. The bottom section allows users to add additional cases either by searching the case repository with keywords (e.g., "brainstorm," as shown in this figure) or by authoring a new case. When adding a case to a policy, users label it to indicate whether they believe the current version of the policy would allow it, disallow it, or whether it is ambiguous how the policy would treat this case. Users can optionally brainstorm with the built-in AI assistant to generate cases that illustrate the policy or identify its potential flaws.



(a) (b) (c)

Figure 5: (a) Once a user adds the case shown in Figure 4 to the policy, it will appear as a card in the related cases section of the policy. The yellow message highlights the misalignment between the label linking the case to the policy (allow) and the majority vote on the case (disallow). (b) Users can click on the card to open a modal with additional details about the case, including other users' reasons for wanting to allow or disallow it. (c) For more in-depth discussions, users can click the "open case" button to visit the case page.

before submitting their changes, users are asked to review the labels of the policy's related cases to see if their policy edit results in any updates to case labels. For example, a case that was previously allowed by the policy might now be disallowed following their edit. After a user updates the labels as needed and submits their edits, the revised policy is made immediately visible to others. If a user's edits resolve a misalignment between the label and people's votes on a case, the alert on that case will disappear.

4.3.2 ***Users are encouraged to be bold in editing.*** The current version of PolicyCraft enables any user to revise policies and labels, and includes a feature to detect editing conflicts in Wikipedia style [46]. Specifically, if a user submits a change to a policy while another user is still editing it, the second user will be asked to review the new changes and decide whether to include them with their own edits before submitting. This design aims to encourage participation (D3), rather than placing too many restrictions on users. To support coordination, the system makes the evolution of a policy transparent [20, 84]: all users can see the full edit history and revert changes as needed.

4.4 Proposing Policies Through Cases

If users notice a gap in current policies, they can create a new policy by clicking the **create policy** button, which takes them to the **creation page**. To encourage users to design and deliberate around policies based on concrete cases (D1), users are reminded on the creation page that policies must include at least one related case in order to eventually be included in the policy finalization stage. On the creation page, users can also choose to create a new case that is not yet related to any policies. Users can first create a case, wait for it to receive votes from others, and then create a policy based on people's perspectives on that case.

4.5 Additional Features to Support Participation and Collaboration

Below, we briefly describe two additional features of PolicyCraft, aimed at facilitating participation and collaboration.

4.5.1 ***Scaffolding policy design with built-in AI assistants.*** To help reduce barriers to participation in policy design (D4), PolicyCraft has three built-in, LLM-based AI assistants that users can optionally use to support the critique, revision, or creation of new policies based on cases. The first AI assistant is available on the page where users edit a policy's related cases. Given a policy as input, this AI assistant can help users brainstorm *case-based critiques* (cases that reveal potential flaws or ambiguities in the current policy) or *illustrative cases* (cases that can help illustrate the application of the policy). The second AI assistant is available on the page where users edit a policy. Given one or more user-selected cases, along with the user's reasons for wanting to allow or disallow those cases, this AI assistant helps users brainstorm ways to *revise* a given policy. The third AI assistant is available on the creation page where users can propose a new policy. Given one or more user-selected cases, along with the user's reasons for wanting to allow or disallow those cases, this AI assistant helps users brainstorm a *new policy*. The design of the AI assistant's conversational flow can be found in Figure 7.

4.5.2 ***Users can participate in discussions, receive notifications, and track their activities within the system.*** To facilitate user collaboration, PolicyCraft includes built-in features for discussion, notifications, and activity tracking (D3). Each policy and case has its own discussion panel where users can start and reply to discussion threads on various topics. We separate discussions about policies and cases to help users clarify if disagreements are about a policy or specific cases (D2). PolicyCraft also has a panel for meta-discussions on the main page (the **about page**) where users can talk and coordinate about general topics that go beyond individual policies and cases. This design is similar to Meta Stack Overflow, where users can have higher-level discussions beyond individual posts [26]. Users receive notifications whenever new discussions are added to threads they have participated in, or when there are changes to policies they are following. Finally, users can keep track of their own activities on a dedicated activity page and quickly revisit the policies, cases, and discussions where they have previously contributed.

4.6 Implementation

PolicyCraft is a full-stack web application that people can set up and invite others to join via a URL. The current implementation of PolicyCraft is built with SvelteKit and uses Firestore databases. On the front end, we use shadcn-svelte components and customize them with Tailwind CSS. On the back end, in addition to the database, we also use Firebase for user authentication, hosting the server, and tracking database changes to send user notifications. Last but not least, we power the AI assistants with Gemini 1.5 Pro. PolicyCraft is open-source and available on GitHub.[5]

5 Field Study

To understand how people use PolicyCraft and how its case-grounded approach shapes collaborative policy development, we conducted an evaluation study in university classes. In this context, we aim to support students in collaboratively designing course policies regarding which uses of generative AI should be allowed in class. Since students are directly impacted by course policies, engaging them in the design process, rather than leaving it solely to instructors, has potential to produce policies that students find more reasonable and legitimate [32, 64, 79]. This process also provides students to learn more about each other's perspectives and build consensus around course practices and policies.

Our evaluation primarily consists of a three-day field study. We ran the study in two classes taught by two of the co-authors who were open to incorporating students' input into their course policies on generative AI. Both courses were electives open to undergraduate and graduate students, focusing on the study and design of human interactions with technologies. In each class, we randomly divided the students into two groups. One group of students collaboratively designed course policies using the **full version of PolicyCraft**. The other group of students used a **baseline version**, which only included the policy-level features (policy creation, editing, and discussion), without any case-level features such as case creation, the case repository, or the AI assistants that scaffold case-grounded policy design. Students in the baseline condition

[5] https://github.com/tskuo/PolicyCraft

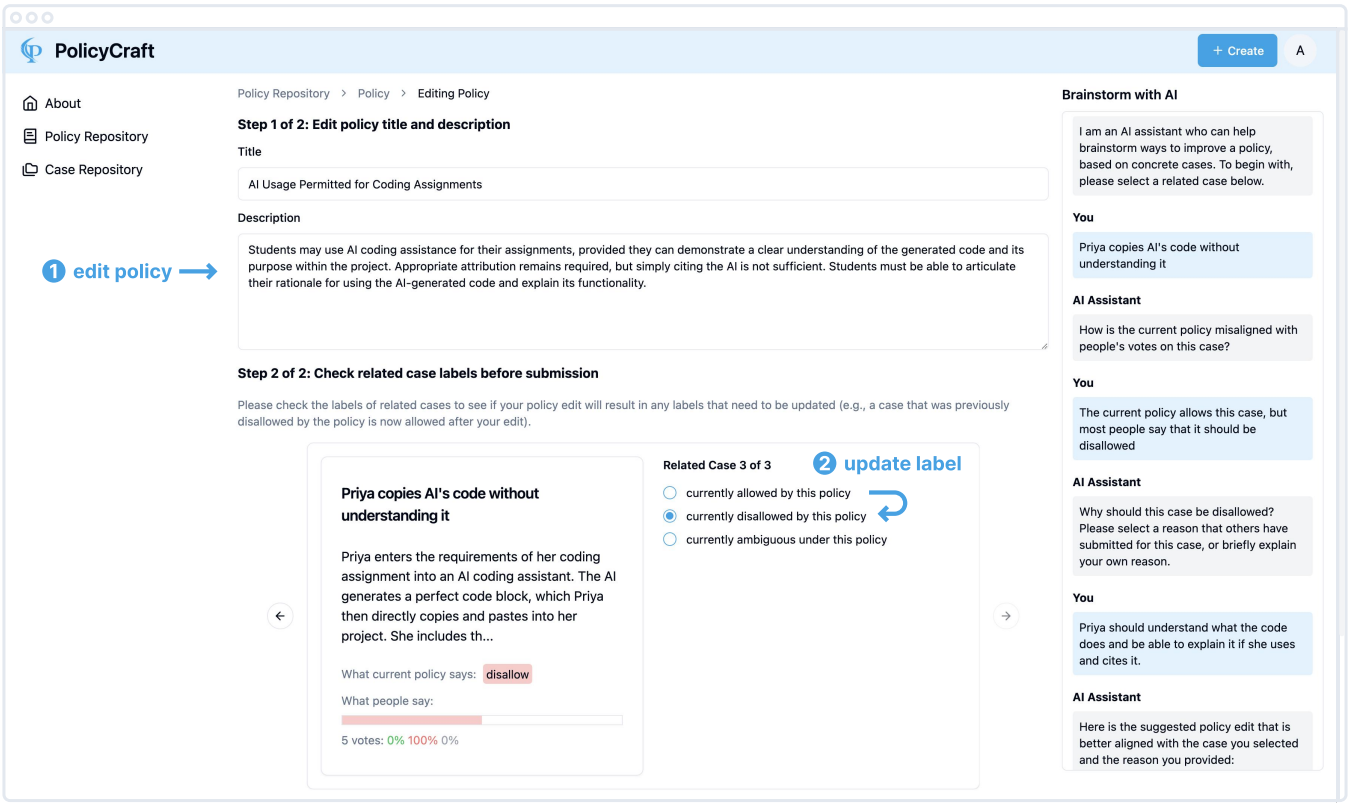


Figure 6: The workflow for editing a policy in PolicyCraft. Users first update the policy's wording and then check if the edit requires updating the labels for any of the related cases. Users may also brainstorm with the built-in AI assistant to improve the policy. For example, the policy description shown in this figure is suggested by the AI assistant based on the user's interactions shown in the right-side panel.

were still able to discuss cases, if they so chose, using the system's discussion features. However, they were not explicitly supported by the system in discussing and using cases to drive iterative policy design. Students in each class designed their course policies independently from the other class, as each class had its unique learning goals. The number of students in each class and condition is shown in Table 1. In addition to the main field study, we conducted two brief surveys to better understand students' perceptions of the process and external raters' opinions on the final policies. All studies were approved by the university's institutional review board (IRB) where studies were conducted.

5.1 Study Procedures

On the first day of the study, students spent 45 minutes onboarding and beginning to use the system during class time. In each class the instructor introduced the study and randomly divided the students into groups, seated in separate areas of the classroom. Students in both groups then individually reviewed onboarding materials illustrating how to use the version of the system they were assigned to use (either the full version of PolicyCraft or the baseline version). Across classes, both versions of the system were initialized with the same set of three initial policies provided by the instructors to kick-start the discussion. In the full version of PolicyCraft each policy included two initial, illustrative cases provided and labeled by the instructors. The initial policies and cases are available in Appendix C. Students were encouraged to use the system for discussions instead of discussing verbally, to leverage the system's collaboration features and avoid cross-group influence. By the end of the first class, all students had begun using the system.

Between the first and second class sessions, which were separated by one day for both participating classes (e.g., Monday and Wednesday), students were asked to use the system during a minimum of three periods: first, as homework after the first class, due by the end of the day; second, anytime during the day between classes; and finally, a third time anytime before the start of the second class. During each period, students were asked to complete at least seven actions, including creating or editing at least one policy. For the full version of PolicyCraft, all actions related to policies, cases, reasons, and discussions counted, except for voting on cases or 'liking' reasons added by others, since these actions required only a single click. For the baseline version, all actions related to policies and discussions counted. We expected students to meet these minimum participation requirements easily. For example, a student using the full version of PolicyCraft could quickly complete three actions by creating a case, explaining the reason behind their vote on that

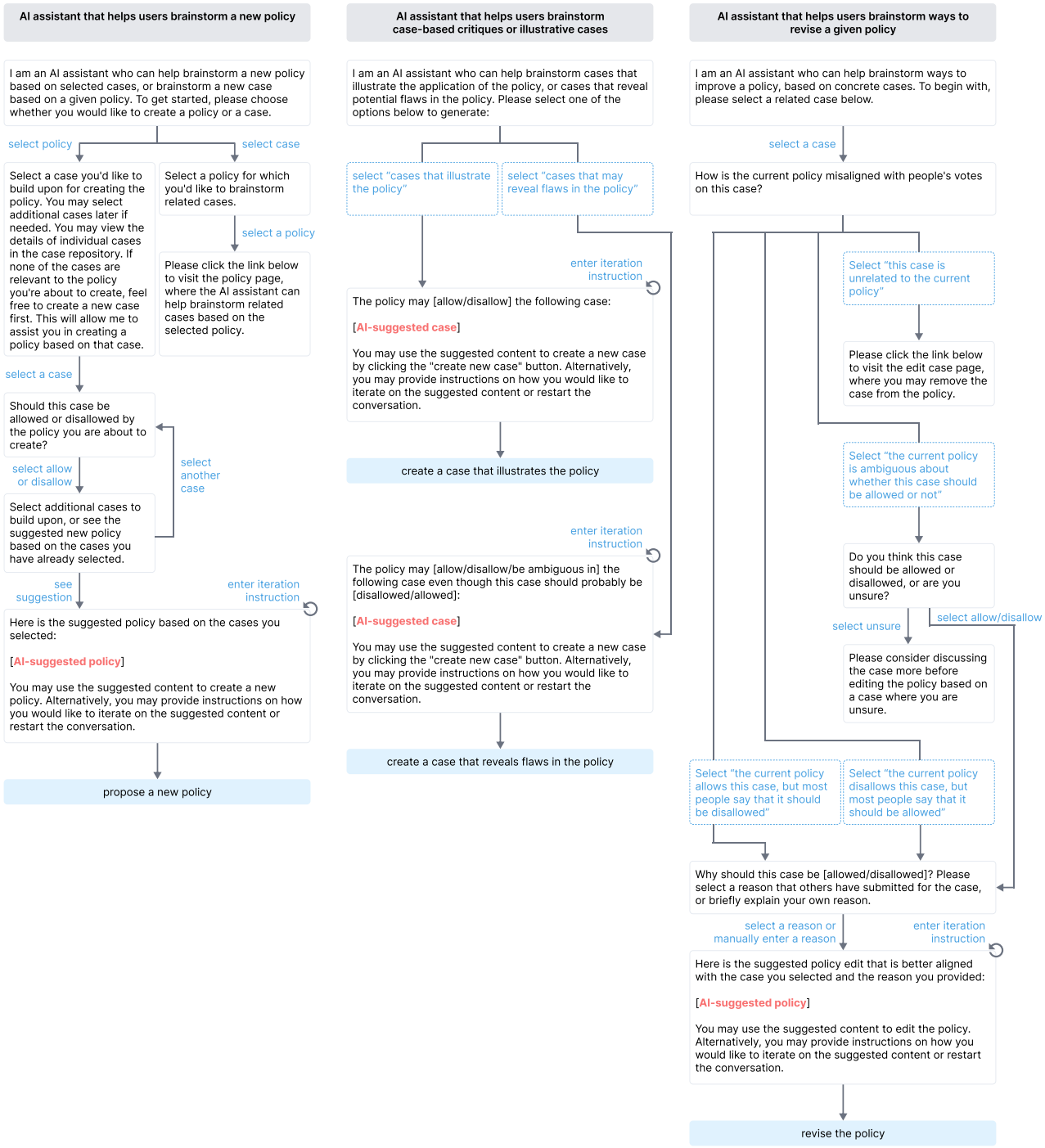


Figure 7: The conversational flows of PolicyCraft's three AI assistants. The blue text represents the user's selection or input, while the black text in boxes shows the AI's response. The red text represents placeholders for AI-suggested policies or cases, based on the information provided by users during the conversation. Users can optionally provide additional instructions to iteratively refine initial AI generations. Users can choose to restart the conversation at any point, but the chat history remains available for their reference. See Appendix B for a more detailed walkthrough of the conversation workflow, including prompts.

case, and adding it to a policy. We set minimal participation requirements to ensure that students would have ample opportunities for interaction across the field study period, while also providing students with flexibility to decide when and how much they want to contribute (cf. [52, 98]). Students were graded on whether they met these minimum participation requirements, but not on the content of their contributions. All students met or exceeded the requirements, knowing that the policies they collaboratively developed would inform the official course policies.

Each time students created or edited a policy, they answered a multiple-choice question about what inspired them to create or edit the policy. By analyzing how often each option was chosen, we could compare inspirations for policy creation and edits in each of our study conditions. Students could choose one or more of the following five options:

1. To address a specific case/scenario that I thought of
2. To address a specific case/scenario that someone else shared
3. To address a general issue that I thought of
4. To address a general issue that someone else shared
5. Other

In the second class session, students within each group voted on which policies should be incorporated into the official course policy using upvotes and downvotes (part of the "finalization" stage of PolicyCraft as shown in Figure 2). Students then individually completed a post-study survey (Appendix D) to understand their experiences with the process. Finally, students participated in a full-class discussion about the resulting policies, facilitated by their instructors. Instructors then used students' feedback to refine the best-supported policies into official course policies on generative AI use for the rest of the semester.

Finally, to understand how people who did not participate in policy development perceived the quality of the resulting policies, we conducted a short online survey after the field study. We recruited university students not involved in the field study as external raters to rate the policies that received majority upvotes from participants in each group. Each rater evaluated two sets of policies—one from the baseline condition and one from the PolicyCraft condition—selected at random from one of the two classes. They were not told how the policy sets were developed. They rated individual policies based on *quality*, *clarity*, and *feasibility* (for implementation in class). They also rated each policy set holistically for *comprehensiveness* and *coherence*. Lastly, they selected which policy set (if any) they felt was higher quality overall, and briefly explained their response. As an incentive for participation, survey participants had the option to enter a raffle for a $100 gift card.

6 Study Results

We present findings from our evaluation study in the following subsections. Section 6.1 presents the resulting policies and analyzes participants' votes on these policies. Our results show that policies developed using PolicyCraft received stronger support and achieved greater consensus among the students who developed them. Section 6.2 explores possible mechanisms to explain this finding. From our post-study survey, we find that students using PolicyCraft found it easier to understand each other's perspectives. We also find evidence that PolicyCraft succeeded in promoting case-grounded, collaborative policy design: Students using PolicyCraft were more likely to iterate on policies based on cases raised and discussed by others. Section 6.3 presents perceptions of the resulting policies from both course instructors and external raters. Overall, instructors and raters found the policies developed with PolicyCraft to be clearer and more nuanced regarding when generative AI use is appropriate, but also less concise and potentially *too* comprehensive for direct use as course policies.

6.1 Resulting Policies and Vote Distributions

Overall, participants using the baseline system proposed a larger number of policies than participants using PolicyCraft. However, in the baseline condition, most policies did not receive majority support in the final voting stage. Table 1 shows the total number of policies created by each group and the number of those policies that received majority votes. To provide a glimpse of the resulting policies, in Table 2 we sample a few policies developed with each system that received a majority of upvotes and policies that did not from their groups. In Appendix E, we include the full set of policies that received majority support from each group within each class, for reference.

6.1.1 ***Policies developed through PolicyCraft received stronger support.*** To understand whether PolicyCraft helped groups collaboratively develop better-supported policies, compared with the baseline system, we analyzed the number of policies that received a majority vote from participants. Specifically, we counted the number of policies where the difference between upvotes and downvotes exceeded half the number of participants involved in their development. As shown in Table 1, 74% of the policies (14 out of 19) developed with PolicyCraft in Class 1 received majority support from participants, compared to just 23% in the baseline condition (7 out of 31). Similarly, 73% of the policies developed with PolicyCraft in Class 2 received majority support from participants, compared to 37% in the baseline condition. These results suggest that participants using PolicyCraft produced better-supported policies overall. In contrast, participants using the baseline system spent more effort creating policies that eventually failed to gain substantial support from the group.

6.1.2 ***Policies developed through PolicyCraft achieved greater consensus during voting.*** In addition to comparing the percentage of policies that received a majority vote, we also wanted to understand whether PolicyCraft helped participants reach greater consensus in their voting across *all* policies. A higher level of consensus on a policy occurs when most participants either upvote or downvote it. In contrast, a lower level of consensus is reflected by roughly equal numbers of upvotes and downvotes. To quantitatively assess the level of consensus on policies, we used Shannon Entropy, a metric from information theory [76] commonly used to measure the degree of consensus in groups [3, 4, 88]. A lower entropy indicates a higher level of consensus. We first computed the entropy for each policy based on its vote distribution:

$$-(p_u \log_2 p_u + p_d \log_2 p_d)$$

where p_u and p_d denote the proportion of upvotes and downvotes a policy received. For example, if a policy receives an equal number of upvotes and downvotes, both p_u and p_d would be 0.5. In this

Table 1: Descriptive statistics of the resulting policies and the votes they received from participants.

	Class 1		Class 2	
	Baseline	PolicyCraft	Baseline	PolicyCraft
Number of participants	22	20	14	12
Number of policies	31	19	19	11
Number of policies with majority upvotes	7	14	7	8
Percentage of policies with majority upvotes	23%	74%	37%	73%
Entropy based on votes (lower indicates higher consensus)	0.64	0.47	0.50	0.27

Table 2: Illustrative examples of policies that received majority upvotes and policies that did not, from Class 1. The number next to each arrow shows how many upvotes or downvotes each policy received.

	Baseline	PolicyCraft
Examples of Policies with Majority Upvotes	**AI as a Tool, Not a Substitute:** Students could be taught to use AI as a resource to enhance their learning, rather than relying on it to do their work for them. AI can be used for tasks such as research, data analysis, and language translation, but it should not replace critical thinking, problem-solving, or creativity. In addition, it would be useful to know which AI tools and prompts were used that helped with the research to give credit to the tool. Votes: 16 ↑ 0 ↓	**AI for Course Understanding:** Students are permitted to utilize AI to enhance their understanding of course material, such as clarifying complex topics or visualizing key concepts. However, all submitted work must reflect the student's own analysis and understanding. While AI tools can provide guidance and support, direct copying or paraphrasing of AI-generated content is strictly prohibited (this includes drawing your comments on readings from any summary/analysis content that the AI provides.) Votes: 18 ↑ 1 ↓
Examples of Policies without Majority Upvotes	**Open AI Conversation History Policy:** Any generative / conversational AI used in the completion of an assignment should have detailed and chronologically marked logs of any exchange with said AI tool. This log (chat history, prompts, individual instructions for the tool, supplemental information...) should be made available to the instructor as part of the hand-in of the assignment if AI-generated content was directly used in an assignment. Votes: 4 ↑ 10 ↓	**Including ChatGPT Chat Log of the Related Usage:** When students used GenAI for this course's learning purposes, they should cite the usage with the chatlog by pasting the chatlog url into the assignment submission to reinforce more self-regulated usage of LLM tools, even if GenAI was just used for brainstorming. When you work is directly a result of LLM tools either in direct idea (including summary or analysis of a reading), you must submit the chatlog url as a citation.) Votes: 4 ↑ 13 ↓

case, the entropy is 1, indicating the *lowest* level of consensus. In contrast, when all votes for a policy are either upvotes or downvotes, the entropy is 0, indicating the *highest* level of consensus.[6] Table 1 shows the mean entropy across policies for each condition.[7] In both classes policies developed using *PolicyCraft* had a lower entropy compared to those developed using the baseline system, indicating a higher level of consensus in their voting on policies.

6.2 Reasons for Greater Policy Support and Consensus

To understand potential mechanisms behind these trends, we examined the results from the post-study survey that students completed at the end of the field study, as well as the multiple-choice question that participants answered each time they edited or created a policy.

6.2.1 ***Participants who used PolicyCraft found it easier to understand why people agreed or disagreed with each other.*** In the post-study survey, participants rated their agreement, on a 7-point scale, with each of five statements about their experiences in collaborative policy design (see Appendix D). As shown in Figure 8, we observe that participants who used PolicyCraft provided significantly higher ratings regarding whether they could *"easily understand why people agree or disagree with each other"*, controlling for between-class differences ($p < 0.01$, see Appendix F.2 for further details on our regression model[8]). This result suggests that participants who used PolicyCraft found it easier to understand each other's perspectives during the process of iteratively developing policies, which may have contributed to greater consensus and support for policies observed during voting. As a participant reported in the post-study survey: *"We can argue over cases and everyone's comment is clearly visible. The voting system provides a clear way for people to share opinion towards a specific case/policy. The voting system along with the comments makes the agreements/disagreements very clear."*

[6]The value at 0 are given by the limit $0 \log 0 := \lim_{x \to 0^+} x \log x = 0$

[7]The distribution of entropies for each condition is provided in Appendix F.1.

[8]In Appendix F.2, we present results from both a linear and an ordinal regression. The appropriateness of using ordinal versus linear regression for the analysis of Likert scale data has been a subject of wide debate, with recent scholarship showing that each approach has complementary benefits and drawbacks [74]. Our reported findings are robust to the choice of ordinal or linear regression.

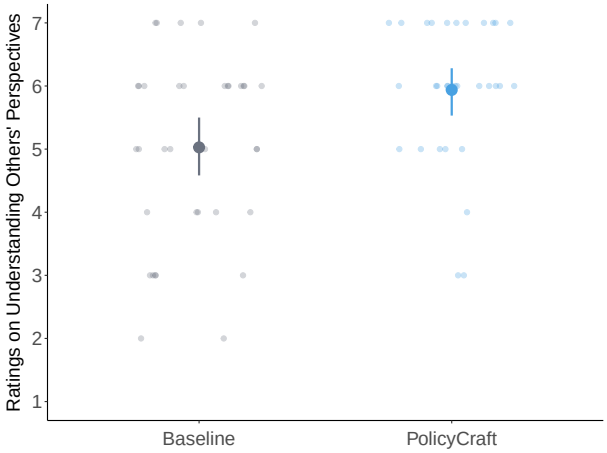


Figure 8: Participants' ratings on whether they could easily understand why people agree or disagree with each other while using the system. There is a significant relationship between the system a participant used and their ratings.

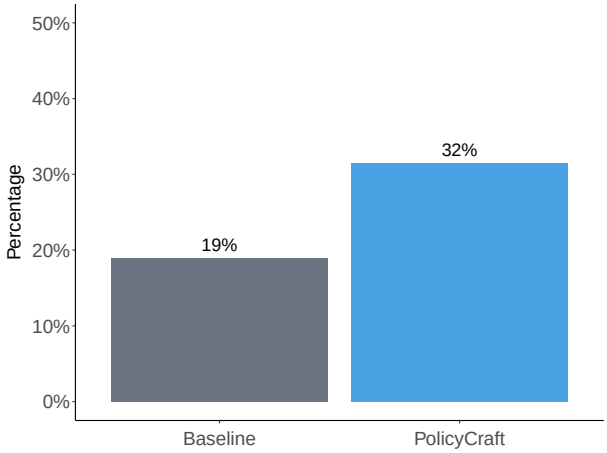


Figure 9: The percentage of policy editing and creation aimed at addressing *specific cases shared by others* during the policy development process. PolicyCraft has a significantly larger proportion than the baseline system.

6.2.2 ***Participants using PolicyCraft were more frequently inspired to edit or create policies based on cases shared by others.*** In the multiple-choice question that participants answered each time they edited or created a policy, participants were asked whether they were inspired to take this action based on a *specific case* or a *general issue*, and whether the case or general issue was something they had *thought of themselves* or that *someone else had shared*. Both PolicyCraft and the baseline system had a similar proportion of policy editing and creation driven by specific cases (58% and 53%, respectively). This result is unsurprising, given that people are known to reason about concrete cases in order to support iterative policy development (see Section 2.3). However, a one-sided hypothesis test reveals that in PolicyCraft, a larger proportion ($p < 0.05$) of policy editing and creation is aimed at addressing specific cases shared by others (32%), compared to the baseline system (19%). This result suggests that PolicyCraft helped participants more often edit and create policies based on cases they collaboratively developed and discussed, with visibility into other participants' perspectives.

Figure 10 illustrates how participants used cases to drive collaborative policy iteration, using a real example from our study. As shown, the initial version of the policy was drafted by a participant based on a case that had received some votes and discussion, indicating mixed views among participants. The initial policy allowed students to use AI for course project brainstorming, so long as the ideas "ultimately come from students themselves". Soon after the policy was created, another participant added a case that they believed should be allowed but which they believed would be disallowed under the current policy (shown in the right side of Figure 10). Other participants voted in agreement. Because the majority vote on this case was misaligned with the policy's label, PolicyCraft highlighted the case with a yellow message noting: *"The policy may need editing to better align with the majority vote on this case."* Finally, one participant revised the initial policy description to address the misalignment. After additional iterations, this policy eventually received majority support. This concrete example shows how PolicyCraft helped participants iterate on policies by discussing cases and building consensus, even when they shared different perspectives.

6.3 External Perceptions of Policies

6.3.1 ***The instructors found the policies comprehensive and creative, but wished students had additional support in synthesizing and generalizing policies.*** To understand the instructors' views on the resulting policies, compared with those they created without students' inputs in previous semesters, we asked them to review the policies and write down their observations and reflections prior to reviewing our study results (cf. [82]).

The instructor of Class 1 found that the policies students collaboratively created *"comprehensively address all aspects relevant to their class experiences, from AI use for conceptual understanding, original work, group work, reading reflections, presentations, coding, and the use of sensitive information."* This instructor also found some policies particularly creative and unexpected, such as one that regulated instructors' behaviors by *"prohibiting AI-generated grading and feedback"* and another meta-policy called the *"three-strike system"* that governed the enforcement of all the policies. This instructor found that *"the PolicyCraft version created a policy set that has a much higher quality than the baseline condition."*

The instructor of Class 2 agreed but felt the final set of policies *"did not quite feel like a finished set of policies yet."* This instructor found that policies developed using PolicyCraft *"reflected more use-case-specific considerations."* While the instructor found this valuable *"because the policies captured really important nuances"* it also led to *"similar policies that should be combined"*. As a result, the instructor found it *"difficult to say which policy set is better overall because one set of policies was more concise and worded more generally, while the other set was more nuanced but also more redundant."*

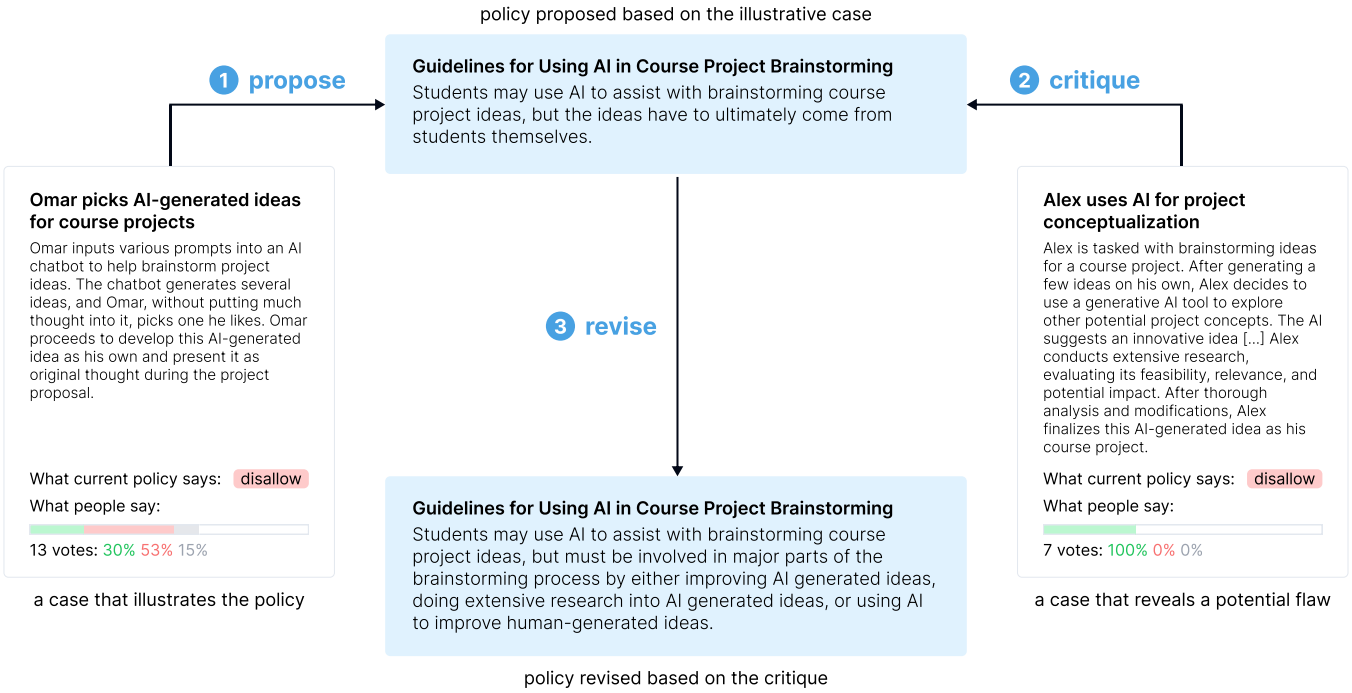


Figure 10: An illustrative example of how participants used cases to iterate on a policy during our field study, showing real policies and cases that our participants developed, voted on, and discussed.

Both instructors wished PolicyCraft had more support for students to synthesize and generalize the policies they developed. For example, the instructor of class two wanted a *"dedicated clustering phase for participants to do some synthesis of their policies"* before voting. In lieu of having this support within the current version of PolicyCraft, each instructor took an initial pass at synthesizing the student-developed policies into final policy sets themselves after the study, and then shared the result with students for any additional feedback.

6.3.2 ***External raters found the policies developed with PolicyCraft clearer on appropriate uses of generative AI, but also less concise and potentially too comprehensive.*** Finally, to understand how students outside of these classes, who did not participate in policy development, perceived the quality of the resulting policies, we analyzed the survey responses from external raters, who evaluated the policies that received a majority vote within each class and condition (without knowledge of how these policy sets were generated). While the ratings did not show a statistically significant difference in external raters' assessments across conditions, feedback from external raters provides insight into the qualitative differences they perceived between the policy set, revealing broad alignment with the course instructors' assessments. For example, a rater who preferred the set of policies developed using PolicyCraft in Class 1 commented that this set of policies *"is clearer about what are appropriate uses of generative AI. [The other set] seems to have a broader interpretation of which uses are okay, leaving it mostly up to interpretation, which creates more of a gray area."* Similarly, another rater noted that the set of policies developed using PolicyCraft: *"breaks down GenAI policies by particular use making it easy to answer question in the form, 'If I am doing X, am I allowed to do Y?' whereas [the other set] comes off more as a list of sentiments about GenAI lightly adapted into an assortment of policies."* In contrast, a rater who preferred the policies developed with the baseline system mentioned: *"I feel [this set] is clearer and easier to implement for the students and instructors."* Another rater also mentioned that the set developed using PolicyCraft seemed: *"a little too comprehensive, to the point where students and instructors may struggle to keep track of what is and isn't allowed."* The raters who evaluated policies in Class 2 provided similar feedback. For example, one rater preferred the PolicyCraft set because *"[the other set] is vague in the exact moments it purports to be specific,"* while another rater preferred the baseline set because they felt the PolicyCraft set was *"harder to implement and less concise."* This overall perception is consistent with the instructors' feedback. In the Discussion section, we discuss how future systems might better support participants in striking a balance between specificity versus generality and conciseness.

7 Discussion

It is crucial that community policies reflect the values and needs of the communities they impact, and that they are viewed as legitimate by community members. In this paper, we present PolicyCraft, a system that supports communities in collaboratively proposing, critiquing, and revising policies through discussion and voting on cases. We conducted a field study across two university courses to understand how people use PolicyCraft in practice. Overall, we

Table 3: Illustrative examples of policies on similar topics developed in each class. Both instructors agreed that the policies developed with PolicyCraft were more grounded in specific use cases (highlighted in bold).

	Baseline	PolicyCraft
Using AI for Programming (Class 1)	*AI for Resolving Coding Errors:* Students should be allowed to use specific AI tools to **fix the coding errors they come across**.	*AI Usage Permitted for Coding Assignments:* Students may use AI to aid in coding assignments, but must use AI to augment their work, not create the solution for them. Students cannot use AI to **create large chunks of code without verifying it themselves**. AI generation of **very broad high-level pseudocode** is permitted, but **not step-by-step pseudocode or detailed lines of code**. AI can be used to **add comments/documentation to already written code** but students should review over them. AI usage must be appropriately attributed.
Acknowledging Use of AI (Class 2)	*Universal AI Attribution Policy:* If AI was used for **a particular assignment**, written notice must be given to the professor using the appropriate technology of submission for that assignment (e.g. comments in one's code if programming, the comment box of a canvas submission, etc.) outlining how AI was used in a particular work. This policy takes precedence over all other policy and is necessary to prevent any legal copyright/cheating issues.	*Usage of GenAI Tools should be Referenced and Cited:* In **assignments or presentations**, students should declare and cite the GenAI tools and prompts that they have used to create content or help that they have received openly. While some cases involve **grammar checks and simple paraphrasing for fluency**, the original idea comes from the student. However, when students use GenAI to **generate code or ideas for responses**, it is essential to add a reference.

found that students using PolicyCraft reached greater consensus and developed course policies with greater community support, compared with those using a baseline system (Section 6.1). Students using PolicyCraft found it easier to understand each other's perspectives, and were more likely to iterate on policies based on concrete cases shared and discussed by others (Section 6.2). Finally, we present external views on the differences between policies developed with PolicyCraft and those created with the baseline system (Section 6.3).

Taken together, while community-external raters' quantitative evaluations did not show a statistically significant difference in the perceived quality of the resulting policies across conditions, their qualitative feedback revealed that policy quality is a complex concept with multiple dimensions shaped by individual preferences. For example, when considering the dimension of policy specificity versus generality, external raters expressed differing preferences that significantly influenced their perception of a policy's quality. Still, even if policies developed using PolicyCraft are not necessarily "higher quality" in a global sense, they received stronger support and consensus from community members who would be impacted by the resulting policies. This indicates that PolicyCraft's case-grounded approach to collaborative policy design can support the development of policies that are better aligned with local community perspectives, and that may be viewed as more legitimate by community members. It is possible that policies received greater community support, in part, simply because participants felt they had influence in shaping the policies. Indeed, this is a well-documented benefit of participation in design. However, given that participants in both the baseline and PolicyCraft conditions were directly engaged in collaboratively shaping policies, such participation effects cannot fully explain the observed advantages of PolicyCraft's case-grounded approach. In this section, we discuss future directions for HCI systems to support collaborative and participatory policy design.

7.1 Balancing Specificity and Generality in Policy Design

Our overall focus in designing PolicyCraft was to support groups in effectively bridging between abstract policies and concrete cases during collaborative policy design. Policy proposals are often high-level and abstract, leaving much open to interpretation, which can make it challenging for groups of people to understand where and why they disagree. By contrast, PolicyCraft assists users in developing policies based on collaborative discussion and consensus-building around specific cases. As mentioned by external raters and course instructors, and illustrated in Table 3, policies developed with PolicyCraft in our study included more case-specific distinctions, making it easier for readers to answer questions like: *"If I am doing X, am I allowed to do Y?"* However, as a consequence, the policies also became less concise and harder for readers to keep track of.

A key insight from our study is that scaffolding for users to *ground* policy development in concrete cases—as provided in the current version of PolicyCraft—needs to be balanced with corresponding scaffolding for users to *abstract* more general policies from these cases. In particular, further research is needed to understand how best to guide participants in striking the right balance between specificity versus generality and conciseness. We expect that the right balance will vary depending on the specific context in which a set of policies will be used. For example, in contexts where only broad guidelines are needed, too much detail can make policies needlessly difficult to implement. It is also worth highlighting the complementary roles of policies and cases, as used in the US legal system, where decisions are based on *both* written laws (statutory law) and past court decisions (case law) [16]. Utilizing both policies and cases for decision-making may help to balance specificity and generality. For example, in the context of PolicyCraft, this could be achieved by presenting more concise versions of the finalized policies *together with* illustrative cases generated by participants,

rather than presenting the policies alone. Future systems should explore mechanisms to scaffold users in developing policies that appropriately balance specificity, generality, and conciseness, tailored to their intended use contexts.

7.2 Supporting Collaborative Review and Synthesis of Policies During Finalization

The current version of PolicyCraft has a "finalization" stage, where participants can upvote or downvote policies and provide reasons for their votes, to support the selection of a final set of policies. However, as the course instructors suggested, it would be helpful for systems like PolicyCraft to explicitly support participants not only in voting on policies during the finalization phase, but also in reviewing the full set of policies and collaboratively synthesizing the policies as needed. For example, future systems could nudge participants to review the entire policy set and discuss whether they want to merge similar policies or split complex policies into multiple simpler ones. This could not only support the removal of redundancies across policies, but could also help participants refine policies to strike a better balance between specificity versus generality and conciseness. In our study, instructors refined and synthesized policies after the voting stage and then shared them with students via Google Docs for additional feedback before finalizing them as the official course policies. Future systems could integrate this process more effectively to better support end-to-end collaborative policy design.

7.3 Making Sense of an Evolving Space of Policies and Cases

One challenge participants in our study faced was navigating and making sense of the space of current policies and cases as their numbers grew and the content of each evolved. Adding more support for review and synthesis of policies during policy finalization could help to address this problem. However, it would also be useful to support users in more effectively making sense of the evolving policy–case space *throughout* the collaborative policy development process. Future systems could draw inspiration from prior research on collective sensemaking by including visualizations that provide an overview of current policies and cases, along with the ability to visually track how the joint space of policies and cases have evolved over time, from an aerial view. In addition, sensemaking might be supported through LLM-based summaries that complement such visualizations by supporting rapid, targeted question-answering [61]. In line with prior work, the usefulness of these visualizations or automated summaries could potentially be supported through a user-driven tagging system that enables users to shape the criteria along which similarity between policies and cases is determined [57, 65, 70, 71, 98].

7.4 Advancing AI Assistants in Policy Design

The current implementation of PolicyCraft includes three built-in, LLM-based AI assistants that help users with policy design, as mentioned in Section 4.5.1. Some participants found the AI assistants helpful for brainstorming cases because *"it helps me think on both side[s], what to allow or disallow, and what should be the case scenarios to accept or reject the policy."* Other participants found revising and creating policies *"straightforward and easy, especially with the help of the embedded AI assistant."* However, some participants preferred to *"discuss the real cases in our lives regarding policy crafting because AI-generated cases are quite similar and less reliable than real cases."* Overall, the AI assistants were not used extensively during our study, with only 26 interactions from the 32 students who had access to the full version of PolicyCraft, across the three-day study period. Future research should further explore how AI assistants can most effectively scaffold collaborative policy design processes, and investigate how different forms of AI-based support may shape both the policy design process and the resulting policies.

7.5 Navigating Power Dynamics in Policy Design and Implementation

PolicyCraft is designed to facilitate a collaborative approach to policy design, potentially involving participation and deliberation among various stakeholder groups within a community. It is important to note that PolicyCraft cannot, on its own, overcome power dynamics that may be present among different stakeholder groups. In communities with distributed power dynamics (e.g., Wikipedia), community members may directly implement the policies they collaboratively develop using PolicyCraft. However, in communities with a hierarchical power structure, approval from community leaders may be necessary to actually implement policies. Nonetheless, we envision that even in such communities, PolicyCraft can be used by community members in a bottom-up fashion to generate policy proposals that have strong community support and consensus. We hope PolicyCraft will serve as a source of inspiration for empowering participation to enable a more democratic approach to community governance [36, 52, 75, 99].

7.6 Supporting Collaborative Policy Design Across a Broader Range of Contexts

PolicyCraft is designed to support policy design across a broad range of community contexts. While the current study focuses on supporting students in collaboratively designing course policies on the use of generative AI, we envision its application in online communities like subreddits to develop content moderation policies [39], or in local communities such as vTaiwan to refine regulations for emerging technologies [33]. Still, some implementations of PolicyCraft will require careful adaptation to align with specific communities' norms. For example, to encourage policy iteration while holding participants accountable for their edits, the current implementation of PolicyCraft adopts Wikipedia's approach by allowing all participants to edit policies while maintaining a transparent edit history [46]. While this approach has shown to be effective in supporting coordination and preventing vandalism in peer production [20, 84] and the current study, a more robust moderation strategy may be required when applying PolicyCraft in different contexts. Future research should investigate which aspects of PolicyCraft's current design need to be adapted for use by other communities and identify which findings from the current study are most transferable across diverse contexts.

8 Conclusion

In this work, we have demonstrated how a system that scaffolds users in grounding policy design in concrete cases can support more effective collaborative policy design. Our findings show that policies developed using PolicyCraft receive stronger support and more consensus from those impacted by the policies, compared to policies developed with a baseline system that did not scaffold their use of concrete cases. Building on this work, future HCI research should explore the design of tools and processes to better support collaborative and participatory policy design. This includes helping users balance policy design trade-offs, make collaborative decisions about how to abstract and generalize from concrete cases, better track and understand the evolving space of policies and cases during collaboration, and more effectively identify areas for policy improvements.

Acknowledgments

The funding for this research was provided by UL Research Institutes through the Center for Advancing Safety of Machine Intelligence, CMU's Block Center for Technology and Society, and Metagov's Grant for Interoperable Deliberative Tools. We thank Dominik Moritz, Jodi Forlizzi, Joseph Seering, Maarten Sap, Motahhare Eslami, Suguru Ishizaki, and Tiffany Chih for their insightful feedback on the system design. We are also grateful to CMU's Eberly Center for their guidance on the study design and to Taiwan's g0v community for their assistance in testing the system prototype. Finally, we thank Isadora Krsek for designing the PolicyCraft logo.

References

[1] Agnar Aamodt and Enric Plaza. 1994. Case-based reasoning: foundational issues, methodological variations, and system approaches. *AI Commun.* 7, 1 (mar 1994), 39–59.

[2] Tanja Aitamurto, Peter G Royal, and Jorge Saldivar. 2023. Disagreement, Agreement, and Elaboration in Crowdsourced Deliberation: Ideation Through Elaborated Perspectives. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 88, 10 pages. https://doi.org/10.1145/3544549.3585708

[3] Yoshio Akiyama, James Nolan, Marjorie Darrah, Mushtaq Abdal Rahem, and Lei Wang. 2016. A method for measuring consensus within groups: An index of disagreement via conditional probability. *Information Sciences* 345 (2016), 116–128.

[4] Iván Aranzales, Ho Fai Chan, Reiner Eichenberger, Rainer Hegselmann, David Stadelmann, and Benno Torgler. 2021. Scientists have favorable opinions on immunity certificates but raise concerns regarding fairness and inequality. *Scientific reports* 11, 1 (2021), 14016.

[5] Sherry R Arnstein. 1969. A ladder of citizen participation. *Journal of the American Institute of planners* 35, 4 (1969), 216–224.

[6] Mariam Asad and Christopher A. Le Dantec. 2015. Illegitimate Civic Participation: Supporting Community Activists on the Ground. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada) *(CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 1694–1703. https://doi.org/10.1145/2675133.2675156

[7] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature* 563, 7729 (2018), 59–64.

[8] Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology* (New York, New York, USA) *(UIST '10)*. Association for Computing Machinery, New York, NY, USA, 313–322. https://doi.org/10.1145/1866029.1866078

[9] Kirsten Boehner and Carl DiSalvo. 2016. Data, Design and Civics: An Exploratory Study of Civic Tech. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2970–2981. https://doi.org/10.1145/2858036.2858326

[10] Brian Butler, Elisabeth Joyce, and Jacqueline Pike. 2008. Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 1101–1110. https://doi.org/10.1145/1357054.1357227

[11] Benjamin N Cardozo and Andrew L Kaufman. 2010. *The Nature of the Judicial Process*. Quid Pro Books.

[12] Alissa Centivany. 2016. Policy as Embedded Generativity: A Case Study of the Emergence and Evolution of HathiTrust. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 926–940. https://doi.org/10.1145/2818048.2820069

[13] Alissa Centivany. 2016. Values, ethics and participatory policymaking in online communities. *Proceedings of the Association for Information Science and Technology* 53, 1 (2016), 1–10.

[14] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 32 (nov 2018), 25 pages. https://doi.org/10.1145/3274301

[15] Andrew Chen. 2021. *The cold start problem*. Harper Business.

[16] Quan Ze Chen and Amy X. Zhang. 2023. Case Law Grounding: Aligning Judgments of Humans and AI on Socially-Constructed Concepts. arXiv:2310.07019 [cs.HC] https://arxiv.org/abs/2310.07019

[17] Quan Ze Chen and Amy X. Zhang. 2023. Judgment Sieve: Reducing Uncertainty in Group Judgments through Interventions Targeting Ambiguity versus Disagreement. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 283 (oct 2023), 26 pages. https://doi.org/10.1145/3610074

[18] Inyoung Cheong, King Xia, K. J. Kevin Feng, Quan Ze Chen, and Amy X. Zhang. 2024. (A)I Am Not a Lawyer, But...: Engaging Legal Experts towards Responsible LLM Policies for Legal Advice. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 2454–2469. https://doi.org/10.1145/3630106.3659048

[19] Eric Corbett, Emily Denton, and Sheena Erete. 2023. Power and Public Participation in AI. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Boston, MA, USA) *(EAAMO '23)*. Association for Computing Machinery, New York, NY, USA, Article 8, 13 pages. https://doi.org/10.1145/3617694.3623228

[20] Laura Dabbish, Colleen Stuart, Jason Tsay, and Jim Herbsleb. 2014. Transparency and coordination in peer production. *arXiv preprint arXiv:1407.0377* (2014).

[21] Ranjana Das, Yen Nee Wong, Rhianne Jones, and Philip JB Jackson. 0. How do we speak about algorithms and algorithmic media futures? Using vignettes and scenarios in a citizen council on data-driven media personalisation. *New Media & Society* 0, 0 (0), 14614448241232589. https://doi.org/10.1177/14614448241232589

[22] Anna De Liddo and Simon Buckingham Shum. 2010. Cohere: A prototype for contested collective intelligence. (2010).

[23] Dmitry Epstein, Cynthia Farina, and Josiah Heidt. 2014. The value of words: Narrative as evidence in policy making. *Evidence & Policy* 10, 2 (2014), 243–258.

[24] Sheena Erete and Jennifer O. Burrell. 2017. Empowered Participation: How Citizens Use Technology in Local Governance. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2307–2319. https://doi.org/10.1145/3025453.3025996

[25] Jenny Fan and Amy X. Zhang. 2020. Digital Juries: A Civics-Oriented Approach to Platform Governance. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376293

[26] Jingchao Fang, Jia-Wei Liang, and Hao-Chuan Wang. 2023. How People Initiate and Respond to Discussions Around Online Community Norms: A Preliminary Analysis on Meta Stack Overflow Discussions. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing* (Minneapolis, MN, USA) *(CSCW '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 221–225. https://doi.org/10.1145/3584931.3606966

[27] K. J. Kevin Feng, Quan Ze Chen, Inyoung Cheong, King Xia, and Amy X. Zhang. 2023. Case Repositories: Towards Case-Based Reasoning for AI Alignment. arXiv:2311.10934 [cs.AI] https://arxiv.org/abs/2311.10934

[28] Casey Fiesler and Brianna Dym. 2020. Moving Across Lands: Online Platform Migration in Fandom Communities. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 42 (may 2020), 25 pages. https://doi.org/10.1145/3392847

[29] Casey Fiesler, Jialun Jiang, Joshua McCann, Kyle Frye, and Jed Brubaker. 2018. Reddit Rules! Characterizing an Ecosystem of Governance. *Proceedings of the International AAAI Conference on Web and Social Media* 12, 1 (Jun. 2018). https://doi.org/10.1609/icwsm.v12i1.15033

[30] Thomas C Grey. 1983. Langdell's orthodoxy. *U. PiTT. l. rev.* 45 (1983), 1.

[31] Aaron L Halfaker, Tzu-Sheng Kuo, Ciell Brusse, Kenneth Holstein, and Haiyi Zhu. 2025. Collective Meaning Cascades but Strange Ducks Swim Upstream: Facilitating Collective Meaning-making through Co-development of AI Models. In *Extended Abstracts of the 2025 CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. https://doi.org/10.1145/3706599.3706683

[32] Gerald F Hess. 2007. Collaborative course design: Not my course, not their course, but our course. *Washburn LJ* 47 (2007), 367.

[33] Yu-Tang Hsiao, Shu-Yang Lin, Audrey Tang, Darshana Narayanan, and Claudina Sarahe. 2018. vTaiwan: An empirical study of open consultation process in Taiwan. *SocArXiv* 4 (2018).

[34] Sohyeon Hwang and Aaron Shaw. 2022. Rules and rule-making in the five largest Wikipedias. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 347–357.

[35] Luca Iandoli, Ivana Quinto, Anna De Liddo, and Simon Buckingham Shum. 2014. Socially augmented argumentation tools: Rationale, design and evaluation of a debate dashboard. *International Journal of Human-Computer Studies* 72, 3 (2014), 298–319.

[36] Lilly C. Irani and M. Six Silberman. 2013. Turkopticon: interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 611–620. https://doi.org/10.1145/2470654.2470742

[37] Steven J. Jackson, Tarleton Gillespie, and Sandy Payette. 2014. The policy knot: re-integrating policy, practice and design in cscw studies of social computing. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Baltimore, Maryland, USA) *(CSCW '14)*. Association for Computing Machinery, New York, NY, USA, 588–602. https://doi.org/10.1145/2531602.2531674

[38] Mahmood Jasim, Enamul Hoque, Ali Sarvghad, and Narges Mahyar. 2021. CommunityPulse: Facilitating Community Input Analysis by Surfacing Hidden Insights, Reflections, and Priorities. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference* (Virtual Event, USA) *(DIS '21)*. Association for Computing Machinery, New York, NY, USA, 846–863. https://doi.org/10.1145/3461778.3462132

[39] Shagun Jhaver, Seth Frey, and Amy X. Zhang. 2023. Decentralizing Platform Power: A Design Space of Multi-Level Governance in Online Social Platforms. *Social Media + Society* 9, 4 (2023), 20563051231207857. https://doi.org/10.1177/20563051231207857 arXiv:https://doi.org/10.1177/20563051231207857

[40] Charles Kiene, Andrés Monroy-Hernández, and Benjamin Mako Hill. 2016. Surviving an "Eternal September": How an Online Community Managed a Surge of Newcomers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 1152–1156. https://doi.org/10.1145/2858036.2858356

[41] Kimon Kieslich, Nicholas Diakopoulos, and Natali Helberger. 2024. Anticipating impacts: using large-scale scenario-writing to explore diverse implications of generative AI in the news environment. *AI and Ethics* (2024), 1–23.

[42] Hyunwoo Kim, Haesoo Kim, Kyung Je Jo, and Juho Kim. 2021. StarryThoughts: Facilitating Diverse Opinion Exploration on Social Issues. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 66 (apr 2021), 29 pages. https://doi.org/10.1145/3449140

[43] Soomin Kim, Jinsu Eun, Joseph Seering, and Joonhwan Lee. 2021. Moderator Chatbot for Deliberative Discussion: Effects of Discussion Structure and Discussant Facilitation. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 87 (apr 2021), 26 pages. https://doi.org/10.1145/3449161

[44] Seyun Kim, Jonathan Ho, Yinan Li, Bonnie Fan, Willa Yunqi Yang, Jessie Ramey, Sarah E. Fox, Haiyi Zhu, John Zimmerman, and Motahhare Eslami. 2024. Integrating Equity in Public Sector Data-Driven Decision Making: Exploring the Desired Futures of Underserved Stakeholders. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 366 (Nov. 2024), 39 pages. https://doi.org/10.1145/3686905

[45] Reuben Kirkham. 2023. (Legal Design) Research through Litigation. arXiv:2303.14336 [cs.HC] https://arxiv.org/abs/2303.14336

[46] Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. 2007. He says, she says: conflict and coordination in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 453–462. https://doi.org/10.1145/1240624.1240698

[47] Andrew Konya, Lisa Schirch, Colin Irwin, and Aviv Ovadya. 2023. Democratic Policy Development using Collective Dialogues and AI. arXiv:2311.02242 [cs.CY] https://arxiv.org/abs/2311.02242

[48] Vinay Koshy, Tanvi Bajpai, Eshwar Chandrasekharan, Hari Sundaram, and Karrie Karahalios. 2023. Measuring User-Moderator Alignment on r/ChangeMyView. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 286 (oct 2023), 36 pages. https://doi.org/10.1145/3610077

[49] Vinay Koshy, Frederick Choi, Yi-Shyuan Chiang, Hari Sundaram, Eshwar Chandrasekharan, and Karrie Karahalios. 2024. Venire: A Machine Learning-Guided Panel Review System for Community Content Moderation. *arXiv preprint arXiv:2410.23448* (2024).

[50] P. M. Krafft, Meg Young, Michael Katell, Jennifer E. Lee, Shankar Narayan, Micah Epstein, Dharma Dailey, Bernease Herman, Aaron Tam, Vivian Guetler, Corinne Bintz, Daniella Raz, Pa Ousman Jobe, Franziska Putz, Brian Robick, and Bissan Barghouti. 2021. An Action-Oriented AI Policy Toolkit for Technology Audits by Community Advocates and Activists. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 772–781. https://doi.org/10.1145/3442188.3445938

[51] Travis Kriplean, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett. 2012. Supporting reflective public thought with considerit. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (Seattle, Washington, USA) *(CSCW '12)*. Association for Computing Machinery, New York, NY, USA, 265–274. https://doi.org/10.1145/2145204.2145249

[52] Tzu-Sheng Kuo, Aaron Lee Halfaker, Zirui Cheng, Jiwoo Kim, Meng-Hsin Wu, Tongshuang Wu, Kenneth Holstein, and Haiyi Zhu. 2024. Wikibench: Community-Driven Data Curation for AI Evaluation on Wikipedia. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 193, 24 pages. https://doi.org/10.1145/3613904.3642278

[53] Tzu-Sheng Kuo, Hong Shen, Jisoo Geum, Nev Jones, Jason I. Hong, Haiyi Zhu, and Kenneth Holstein. 2023. Understanding Frontline Workers' and Unhoused Individuals' Perspectives on AI Used in Homeless Services. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 860, 17 pages. https://doi.org/10.1145/3544548.3580882

[54] Jonathan Lazar, Julio Abascal, Janet Davis, Vanessa Evers, Jan Gulliksen, Joaquim Jorge, Tom McEwan, Fabio Paternò, Hans Persson, Raquel Prates, Hans von Axelson, Marco Winckler, and Volker Wulf. 2012. HCI public policy activities in 2012: a 10-country discussion. *Interactions* 19, 3 (may 2012), 78–81. https://doi.org/10.1145/2168931.2168947

[55] Pascale Lehoux, Fiona Alice Miller, and Bryn Williams-Jones. 2020. Anticipatory governance and moral imagination: Methodological insights from a scenario-based public deliberation study. *Technological Forecasting and Social Change* 151 (2020), 119800.

[56] Greg Little, Lydia B. Chilton, Max Goldman, and Robert C. Miller. 2010. Exploring iterative and parallel human computation processes. In *Proceedings of the ACM SIGKDD Workshop on Human Computation* (Washington DC) *(HCOMP '10)*. Association for Computing Machinery, New York, NY, USA, 68–76. https://doi.org/10.1145/1837885.1837907

[57] Michael Xieyang Liu, Tongshuang Wu, Tianying Chen, Franklin Mingzhe Li, Aniket Kittur, and Brad A Myers. 2023. Selenite: Scaffolding decision making with comprehensive overviews elicited from large language models. *arXiv preprint arXiv:2310.02161* (2023).

[58] Weichen Liu, Sijia Xiao, Jacob T. Browne, Ming Yang, and Steven P. Dow. 2018. ConsensUs: Supporting Multi-Criteria Group Decisions by Visualizing Points of Disagreement. *Trans. Soc. Comput.* 1, 1, Article 4 (jan 2018), 26 pages. https://doi.org/10.1145/3159649

[59] Theodore J. Lowi. 1972. Four Systems of Policy, Politics, and Choice. *Public Administration Review* 32, 4 (1972), 298–310. http://www.jstor.org/stable/974990

[60] Narges Mahyar, Michael R. James, Michelle M. Ng, Reginald A. Wu, and Steven P. Dow. 2018. CommunityCrit: Inviting the Public to Improve and Evaluate Urban Design Ideas through Micro-Activities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173769

[61] Bruno Marnette and Colleen McKenzie. [n. d.]. *Talk to the City: an open-source AI tool for scaling deliberation.* Retrieved August 23, 2024 from https://ai.objectives.institute/blog/talk-to-the-city-an-open-source-ai-tool-to-scale-deliberation

[62] Mithical. 2023. *Moderation Strike: Stack Overflow, Inc. cannot consistently ignore, mistreat, and malign its volunteers.* Retrieved August 19, 2024 from https://meta.stackexchange.com/questions/389811/moderation-strike-stack-overflow-inc-cannot-consistently-ignore-mistreat-an

[63] Albert C Molewijk, Tineke Abma, Margreet Stolper, and Guy Widdershoven. 2008. Teaching ethics in the clinic. The theory and practice of moral case deliberation. *Journal of Medical Ethics* 34, 2 (2008), 120–124.

[64] Isabel Moreno-Lopez. 2005. Sharing power with students: The critical language classroom. *Radical Pedagogy* 7, 2 (2005), 23–49.

[65] Meredith Ringel Morris, Jarrod Lombardo, and Daniel Wigdor. 2010. WeSearch: supporting collaborative search and sensemaking on a tabletop display. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work* (Savannah, Georgia, USA) *(CSCW '10)*. Association for Computing Machinery, New York, NY, USA, 401–410. https://doi.org/10.1145/1718918.1718987

[66] Priyanka Nanayakkara, Nicholas Diakopoulos, and Jessica Hullman. 2020. Anticipatory ethics and the role of uncertainty. *arXiv preprint arXiv:2011.13170* (2020).

[67] Blair Nonnecke and Jenny Preece. 2000. Lurker demographics: counting the silent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing*

Systems (The Hague, The Netherlands) *(CHI '00)*. Association for Computing Machinery, New York, NY, USA, 73–80. https://doi.org/10.1145/332040.332409

[68] Elinor Ostrom. 1990. *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press.

[69] Joonsuk Park, Cheryl Blake, and Claire Cardie. 2015. Toward machine-assisted participation in erulemaking: An argumentation model of evaluability. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*. 206–210.

[70] Sharoda A. Paul and Meredith Ringel Morris. 2009. CoSense: enhancing sensemaking for collaborative web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. Association for Computing Machinery, New York, NY, USA, 1771–1780. https://doi.org/10.1145/1518701.1518974

[71] Sharoda A. Paul and Madhu C. Reddy. 2010. Understanding together: sensemaking in collaborative information seeking. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work* (Savannah, Georgia, USA) *(CSCW '10)*. Association for Computing Machinery, New York, NY, USA, 321–330. https://doi.org/10.1145/1718918.1718976

[72] Cynthia R. Farina, Dmitry Epstein, Josiah B. Heidt, and Mary J. Newhart. 2013. Regulation Room: Getting "more, better" civic participation in complex government policymaking. *Transforming Government: People, Process and Policy* 7, 4 (2013), 501–516.

[73] Brandon Reynante, Steven P. Dow, and Narges Mahyar. 2021. A Framework for Open Civic Design: Integrating Public Participation, Crowdsourcing, and Design Thinking. *Digit. Gov.: Res. Pract.* 2, 4, Article 31 (dec 2021), 22 pages. https://doi.org/10.1145/3487607

[74] Alexander Robitzsch. 2020. Why ordinal variables can (almost) always be treated as continuous variables: Clarifying assumptions of robust continuous and ordinal factor analysis estimation methods. In *Frontiers in education*, Vol. 5. Frontiers Media SA, 589965.

[75] Niloufar Salehi, Lilly C. Irani, Michael S. Bernstein, Ali Alkhatib, Eva Ogbe, Kristy Milland, and Clickhappier. 2015. We Are Dynamo: Overcoming Stalling and Friction in Collective Action for Crowd Workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1621–1630. https://doi.org/10.1145/2702123.2702508

[76] C. E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27, 3 (1948), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

[77] Aaron Shaw, Haoqi Zhang, Andrés Monroy-Hernández, Sean Munson, Benjamin Mako Hill, Elizabeth Gerber, Peter Kinnaird, and Patrick Minder. 2014. Computer supported collective action. *Interactions* 21, 2 (mar 2014), 74–77. https://doi.org/10.1145/2576875

[78] Joongi Shin, Michael A. Hedderich, AndréS Lucero, and Antti Oulasvirta. 2022. Chatbots Facilitating Consensus-Building in Asynchronous Co-Design. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 78, 13 pages. https://doi.org/10.1145/3526113.3545671

[79] Ira Shor. 1996. *When students have power: Negotiating authority in a critical pedagogy*. University of Chicago Press.

[80] Divya Siddarth, Daron Acemoglu, Danielle Allen, Kate Crawford, James Evans, Michael Jordan, and E Weyl. 2021. How AI fails us. *arXiv preprint arXiv:2201.04200* (2021).

[81] Christopher Small, Michael Bjorkegren, Timo Erkkilä, Lynette Shaw, and Colin Megill. 2021. Polis: Scaling deliberation by mapping high dimensional opinion spaces. *Recerca: revista de pensament i anàlisi* 26, 2 (2021).

[82] C. Estelle Smith, Bowen Yu, Anjali Srivastava, Aaron Halfaker, Loren Terveen, and Haiyi Zhu. 2020. Keeping Community in the Loop: Understanding Wikipedia Stakeholder Values for Machine Learning-Based Systems. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376783

[83] Anne Spaa, Abigail Durrant, Chris Elsden, and John Vines. 2019. Understanding the Boundaries between Policymaking and HCI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300314

[84] H. Colleen Stuart, Laura Dabbish, Sara Kiesler, Peter Kinnaird, and Ruogu Kang. 2012. Social transparency in networked information exchange: a theoretical framework. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (Seattle, Washington, USA) *(CSCW '12)*. Association for Computing Machinery, New York, NY, USA, 451–460. https://doi.org/10.1145/2145204.2145275

[85] Mei Tan and Hari Subramonyam. 2024. More than Model Documentation: Uncovering Teachers' Bespoke Information Needs for Informed Classroom Integration of ChatGPT. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 269, 19 pages. https://doi.org/10.1145/3613904.3642592

[86] Udayan Tandon, Vera Khovanskaya, Enrique Arcilla, Mikaiil Haji Hussein, Peter Zschiesche, and Lilly Irani. 2022. Hostile Ecologies: Navigating the Barriers to Community-Led Innovation. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 443 (nov 2022), 26 pages. https://doi.org/10.1145/3555544

[87] Ningjing Tang, Jiayin Zhi, Tzu-Sheng Kuo, Calla Kainaroi, Jeremy J. Northup, Kenneth Holstein, Haiyi Zhu, Hoda Heidari, and Hong Shen. 2024. AI Failure Cards: Understanding and Supporting Grassroots Efforts to Mitigate AI Failures in Homeless Services. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 713–732. https://doi.org/10.1145/3630106.3658935

[88] JM Tapia, Francisco Chiclana, Maria José del Moral, and Enrique Herrera-Viedma. 2022. Entropy Based Approach to Measuring Consensus in Group Decision-Making Problems. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 409–415.

[89] Judith Jarvis Thomson. 1984. The Trolley Problem. *Yale Law Journal* 94 (1984), 1395.

[90] Fernanda B. Viégas, Martin Wattenberg, and Matthew M. McKeon. 2007. The hidden order of wikipedia. In *Proceedings of the 2nd International Conference on Online Communities and Social Computing* (Beijing, China) *(OCSC'07)*. Springer-Verlag, Berlin, Heidelberg, 445–454.

[91] Leijie Wang, Nicholas Vincent, Julija Rukanskaitundefined, and Amy Xian Zhang. 2024. Pika: Empowering Non-Programmers to Author Executable Governance Policies in Online Communities. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 925, 18 pages. https://doi.org/10.1145/3613904.3642012

[92] Galen Weld, Amy X. Zhang, and Tim Althoff. 2024. Making Online Communities 'Better': A Taxonomy of Community Values on Reddit. *Proceedings of the International AAAI Conference on Web and Social Media* 18, 1 (May 2024), 1611–1633. https://doi.org/10.1609/icwsm.v18i1.31413

[93] Cedric Deslandes Whitney, Teresa Naval, Elizabeth Quepons, Simrandeep Singh, Steven R Rick, and Lilly Irani. 2021. HCI Tactics for Politics from Below: Meeting the Challenges of Smart Cities. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 297, 15 pages. https://doi.org/10.1145/3411764.3445314

[94] David Wright, Bernd Stahl, and Tally Hatzakis. 2020. Policy scenarios as an instrument for policymakers. *Technological Forecasting and Social Change* 154 (2020), 119972.

[95] Yuxi Wu, W. Keith Edwards, and Sauvik Das. 2022. "A Reasonable Thing to Ask For": Towards a Unified Voice in Privacy Collective Action. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 32, 17 pages. https://doi.org/10.1145/3491102.3517467

[96] Qian Yang, Richmond Y. Wong, Steven Jackson, Sabine Junginger, Margaret D. Hagan, Thomas Gilbert, and John Zimmerman. 2024. The Future of HCI-Policy Collaboration. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 820, 15 pages. https://doi.org/10.1145/3613904.3642771

[97] Angie Zhang, Rocita Rana, Alexander Boltz, Veena Dubal, and Min Kyung Lee. 2024. Data Probes as Boundary Objects for Technology Policy Design: Demystifying Technology for Policymakers and Aligning Stakeholder Objectives in Rideshare Gig Work. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 388, 21 pages. https://doi.org/10.1145/3613904.3642000

[98] Amy X. Zhang and Justin Cranshaw. 2018. Making Sense of Group Chat through Collaborative Tagging and Summarization. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 196 (nov 2018), 27 pages. https://doi.org/10.1145/3274465

[99] Amy X. Zhang, Grant Hugh, and Michael S. Bernstein. 2020. PolicyKit: Building Governance in Online Communities. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 365–378. https://doi.org/10.1145/3379337.3415858

[100] Jonathan L Zittrain. 2019. Three eras of digital governance. *Available at SSRN 3458435* (2019).

[101] Ethan Zuckerman and Chand Rajendra-Nicolucci. 2023. From Community Governance to Customer Service and Back Again: Re-Examining Pre-Web Models of Online Governance to Address Platforms' Crisis of Legitimacy. *Social Media + Society* 9, 3 (2023), 20563051231196864. https://doi.org/10.1177/20563051231196864 arXiv:https://doi.org/10.1177/20563051231196864

A Interview Topics

As mentioned in Section 3, we conducted semi-structured interviews with community organizers from different contexts to understand whether and how the design goals we had derived from prior research literature aligned with real-world needs across a range of community contexts. In each interview, we spoke with community organizers, grounding our conversation in actual past experiences where the development of community policies was needed. Below are the discussion topics used to guide our semi-structure interviews:

- **Context**: What is the community context?
- **Roles**: What are the interviewees' roles and responsibilities as community organizers within their community?
- **Needs**: What is (or was) the reason a policy needed to be proposed or implemented in their community?
- **Processes**: What does the current process for policy development look like in their community? What do they envision as the *ideal* process for the development of community policies?
- **Challenges**: What challenges do their communities currently encounter, (or what challenges do they foresee) in the policy development process?

B AI Assistant Details

As mentioned in Section 4.5.1, PolicyCraft has three built-in LLM-based AI assistants that users can optionally use to create, critique, or revise policies based on cases. This section provides the prompts and more details for these AI assistants. Note that the current implementation of the AI assistants, including the prompts and conversation flows, is not intended to be optimal but serves as a demonstration of how LLMs can be integrated into PolicyCraft's case-grounded deliberation approach to policy design. The current work aims to open up a research space for future studies to further develop LLM-based AI assistants that support policy design through the systematic use of cases.

B.1 AI Assistant for Policy Creation

The AI assistant on the creation page can help users brainstorm new policies based on selected cases. Its conversation flow is shown in the left column of Figure 7. Once a user chooses to create a policy, they can select one or more cases from the case repository and specify whether each case should be allowed or disallowed by the policy they are about to create. After the user finishes selecting cases, PolicyCraft will feed the following prompt to an LLM:

> `You are a helpful assistant focusing on supporting users in creating a new policy. In a few sentences, propose a policy that meets the following criteria. 1. The policy should <allow|disallow> the following scenario: [selected case 1]. 2. The policy should <allow|disallow> the following scenario: [selected case 2]. [...]`

The LLM-generated policy will replace the red text in the left column of Figure 7 and be presented to the user, who can then provide additional instructions to refine the policy further. PolicyCraft will feed the current policy and user instructions to an LLM using the following prompt to generate a refined policy.

> `You are a helpful assistant focusing on supporting users in editing the following policy: [current policy]. In a few sentences, slightly revise the policy without significant changes based on the following instructions: [user instruction].`

When the user is satisfied with the policy, whether refined manually or with the help of the AI assistant, they can proceed to create a new policy.

B.2 AI Assistant for Policy Critique

The AI assistant on the page for editing cases related to a given policy (see Figure 4) can help users brainstorm cases that illustrate or reveal flaws in the policy. Its conversation flow is shown in the middle column of Figure 7.

When the user chooses to create a case that illustrates the policy, PolicyCraft feeds the following prompt to an LLM:

> `You are a helpful assistant focusing on supporting users' reflections on context policies. Here is an overview of the context. [context details]. In a few sentences, provide an example scenario of a character in this context where the character does something and <abides by|violates> the following policy: [policy description].`

The underlined content and context details are configured by community organizers during the initialization of PolicyCraft, based on their specific context. For example, in our study, the context is "course", the character is "student", and does something is "uses AI". The system randomly generates cases that either abide by or violate the policy to ensure balanced coverage.

Similarly, when the user chooses to create a case that reveals flaws in the policy, PolicyCraft feeds the following prompt to an LLM:

> `You are a helpful assistant focusing on supporting users' reflections on context policies. Here is an overview of the context. [context details]. In a few sentences, provide an example scenario of a character in this context where < the character technically abides by the following policy but undermines the policy's intent | the character technically violates the following policy despite genuinely trying to comply | it is unclear whether the character violates the following policy or not >: [policy description].`

The system will randomly generate cases using one of the three prompt templates within the angle brackets to encourage exploration of potential flaws.

The LLM-generated cases will replace the red text in the middle column of Figure 7 and be presented to the user, who can then provide further instructions to refine the case. PolicyCraft will feed the current case and user instructions to an LLM using the following prompt to generate a refined case:

> `You are a helpful assistant focusing on supporting users in editing the following case: [current case].`

```
In a few sentences, slightly revise the case without
significant changes based on the following instructions:
[user instruction].
```

When the user is satisfied with the case, whether refined manually or with the help of the AI assistant, they can proceed to create a new case.

B.3 AI Assistant for Policy Revision

The AI assistant on the page for editing a policy (see Figure 6) can help users brainstorm ways to revise the policy based on a selected case. Its conversation flow is shown in the right column of Figure 7. The AI assistant will ask users to check whether the selected case is unrelated to the policy or inherently ambiguous. In such circumstances, the AI assistant will present the user with appropriate action options, rather than using the case to drive policy revision. When the policy allows or disallows the case, but most people say the opposite, the AI assistant will ask the user to provide a reason by selecting from the reasons submitted by others or manually entering their own. Based on the selected case and the provided reason, PolicyCraft will feed the following prompt to an LLM to generate a suggested policy revision:

```
You are a helpful assistant focusing on supporting
users' revision of the following policy: [policy
description]. In a few sentences, slightly revise
the policy without significant changes so that
the policy <allows|disallows> the following scenario:
[selected case]. Here is the reason why the policy
should <allow|disallow> the scenario: [provided
reason].
```

The LLM-generated policy will replace the red text in the right column of Figure 7 and be presented to the user, who can then provide additional instructions to refine the policy further. PolicyCraft will feed the current policy suggestion and user instructions to an LLM using the following prompt to generate a refined policy:

```
You are a helpful assistant focusing on supporting
users in editing the following policy: [policy
description]. In a few sentences, slightly revise
the policy without significant changes based on
the following instructions: [user instruction].
```

When the user is satisfied with the suggested policy, whether refined manually or with the help of the AI assistant, they can proceed to revise the policy.

C Initial Policies and Cases

At the beginning of the field study, the instructors provided the following three initial policies to kick-start the discussion. These policies were carefully chosen to represent different types of class activities and varying levels of restriction on the use of generative AI, ranging from a complete prohibition to unrestricted use. In the full version of PolicyCraft, each policy included two initial, illustrative cases provided and labeled by the instructors. These seed policies and cases serve to minimize the cold-start problem and establish norms around aspects such as formatting and level of abstraction.

- **Prohibition of AI for Reading Responses**: Absolutely no use of AI is allowed for writing reading responses.
  - *Lukas uses AI to summarize key points from papers*: Lukas uses an AI chatbot to summarize the key points of a dense research paper from the week's reading assignment. He then uses these AI-generated points to form the bulk of his reading response, passing off the AI's analysis as his own original thoughts and reflections. (Label: **disallowed** by the policy)
  - *Ding asks AI to explain complex topics*: Ding is struggling to understand a complex topic presented in the week's readings. She turns to an AI-powered study tool, like a chatbot tutor, to explain the topic in simpler terms. She then uses the chatbot's explanation to help her formulate her reading response. Ding does not directly copy the chatbot's words but only uses its insights as a guide. (Label: **disallowed** by the policy)

- **AI Usage Permitted for Coding Assignments**: Students may freely use AI for coding assignments with appropriate attribution.
  - *Mark submits AI's code as his own*: Mark put his project requirements into an AI code generator. The AI spit back a flawlessly functioning, well-documented program. Mark put his name on the AI's work and hit submit for the assignment. (Label: **disallowed** by the policy)
  - *Priya copies AI's code without understanding it*: Priya enters the requirements of her coding assignment into an AI coding assistant. The AI generates a perfect code block, which Priya then directly copies and pastes into her project. She includes the comment: "Used an AI assistant for help with this part." While Priya attempted to acknowledge the AI's assistance, when the instructor asked later, she was unable to explain what purpose that code was meant to serve, or her rationale for including it. (Label: **allowed** by the policy)

- **Guidelines for Using AI in Course Project Brainstorming**: Students may use AI to assist with brainstorming course project ideas, but the ideas have to ultimately come from students themselves.
  - *Omar picks AI-generated ideas for course projects*: Omar inputs various prompts into an AI chatbot to help brainstorm project ideas. The chatbot generates several ideas, and Omar, without putting much thought into it, picks one he likes. Omar proceeds to develop this AI-generated idea as his own and present it as original thought during the project proposal. (Label: **disallowed** by the policy)
  - *Emily uses AI to revamp her discarded ideas*: Emily had tried brainstorming course project ideas with her group, but it had yielded no good ideas. So Emily fed an AI chatbot her group's "discarded ideas" and the assignment description. Within seconds, the chatbot generated a great idea, which the group decided to use. In this case, the "discarded" ideas that were fed to the AI chatbot technically did come from students themselves, but the final idea was AI-generated. (Label: **ambiguous** under the policy)

D Post-Study Survey

At the end of the field study, students completed a post-study survey where they rated their agreement or disagreement with the following five statements on a scale of 1 to 7, and briefly explained their reasoning.

- Overall, I can easily **collaborate with others** in policy development using the system.
- I can easily **identify potential flaws** in a policy using the system.
- I can easily **revise or create policies** to **address potential flaws** using the system.
- I can easily **understand why people agree or disagree** with each other when using the system.
- I feel **confident and comfortable to contribute** to policy development using the system.

E Resulting Policies

Here are the full sets of policies that received majority support from each group within each class:

- **Class 1 – PolicyCraft condition** (14 policies, excluding 5 without majority vote):
  - **AI Use for Conceptual Understanding**: Student is allowed to use AI for understanding different concepts, or asking for resources to understand a concept in the relevant domain.
  - **Use of Sensitive Information for AI**: Students should not enter sensitive information in their prompts/entries of GenAI models. Sensitive information includes personal identifiable information (ex. school ID, email addresses), copyrighted information from the course, and other confidential information (ex. OpenAI credentials, Azure credentials).
  - **AI for Course Understanding**: Students are permitted to utilize AI to enhance their understanding of course material, such as clarifying complex topics or visualizing key concepts. However, all submitted work must reflect the student's own analysis and understanding. While AI tools can provide guidance and support, direct copying or paraphrasing of AI-generated content is strictly prohibited (this includes drawing your comments on readings from any summary/analysis content that the AI provides.
  - **AI Guidelines for Original Work**: AI-powered tools should be used to support learning, brainstorming, and other academic activities, but they should not replace original work.
  - **AI in Collaborative Group Work**: Students may use generative AI tools to assist with tasks in collaborative group projects, such as idea generation, task delegation, and content creation. However, all group members must be actively involved in the process, and the AI's contributions should be transparently discussed among the team. The final project must reflect the collective effort and understanding of the group, with AI being a supportive tool rather than the primary contributor.
  - **AI for Presentation Preparation**: AI should be allowed to be used for Final Project Presentation as well as other presentations needed to be prepared throughout the course. It may be employed for augmentation such as grammar/spell checking, brainstorming, and template suggestions. However, students cannot directly use AI-generated text, images, or any other content in their presentations. (They can if cited)
  - **AI Citation**: If AI contributes at all towards materials created by the student (ex: reading notes, projects, code, presentations), then the student must acknowledge that they used AI in that assignment/submission.
  - **AI Usage for Grammar Checks and Better Writing**: AI can be used to refine writing to remove grammatical errors, spelling errors etc without changing the actual content of the text.
  - **AI Usage Permitted for Coding Assignments**: Students may use AI to aid in coding assignments, but must use AI to augment their work, not create the solution for them. Students cannot use AI to create large chunks of code without verifying it themselves. AI generation of very broad high-level pseudocode is permitted, but not step-by-step pseudocode or detailed lines of code. AI can be used to add comments/documentation to already written code but students should review over them. AI usage must be appropriately attributed.
  - **Prohibition of AI for Reading Responses**: It is not permitted to use AI to summarize or generate answers for reading responses. However, the usage of AI-powered tools is permitted to edit work (syntax, spelling/grammar checks, translate), or provide explanations for readings for the purpose of understanding the text after reading on one's own. Any usage of AI tools should be cited.
  - **Prohibiting AI-Generated Grading and Feedback**: To ensure the quality and authenticity of student assessments, AI tools may not be used to generate generic grading or feedback unless the assessments are graded based on completion. This policy aims to maintain the integrity of the academic process and provide students with personalized and meaningful evaluations.
  - **AI Use in Debugging**: Students can use AI to help with debugging as long as they have considered the code themselves already and understand the small revisions made.
  - **A 3-Strike System**: Under this policy students who violate established AI usage guidelines will receive progressive consequences, starting with a 50% reduction in assignment score and escalating to more severe penalties like 0 grades, disciplinary referrals, and loss of privileges. (We may need to adjust the specific consequences and procedures based on individual circumstances and perspectives).
  - **Guidelines for Using AI in Course Project Brainstorming**: Students may use AI to assist with brainstorming course project ideas, but must be involved in major parts of the brainstorming process by either improving AI generated ideas, doing extensive research into AI generated ideas, or using AI to improve human-generated ideas.
- **Class 1 – baseline condition** (7 policies, excluding 24 without majority vote):
  - **AI for Resolving Coding Errors**: Students should be allowed to use specific AI tools to fix the coding errors they come across.
  - **AI as a Tool, Not a Substitute**: Students could be taught to use AI as a resource to enhance their learning, rather than relying on it to do their work for them. AI can be used for tasks such as research, data analysis, and language translation, but it should not replace critical thinking, problem-solving, or creativity. In addition, it would be useful to know which AI tools and prompts were used that helped with the research to give credit to the tool.
  - **If You Find Cool Gen AI Application(s), Share Your Favorites with the Class?**: If a student discovers a particularly valuable or interesting application, they are encouraged to share it with the class to inspire and inform their peers. Sharing Platform: We can crearte a Miro Board Sharing Platform. Students create a card on Miro with the AI tool's name, a link, and a brief reason for sharing. Tools can be categorized by purpose (e.g., image generation, text processing). Students can explore and vote on tools at any time. Students can share directly on Miro without taking class time.
  - **Combating Hallucinations**: If using AI as a vehicle for information, we must ensure it's correct so I suggest requiring that people find sources to back info found by chatbots just to make sure information is up to date and right. This also helps make sure that people are still having to do research and not just using whatever the chatbot puts out. Students need to show the attempt of avoiding hallucination for the topics they are not familiar with by: 1) Asking the AI to provide reference for the conclusion it made in the answers 2) Adding necessary requirements in the prompt engineering, e.g. "only answer the questions if you are 100% confident, or else return I don't know" 3) Show extra effort in doing research out of AI for validation (like providing links for reference) 4) provide an explanation of possible sources of hallucination or limitation in the answers, and propose alternative solutions
  - **Course Instructor's AI Acceptance Rules**: After finalization of policies, instructors need to provide clear guidance on what are the acceptable uses of AI and mention if any specific AI use cases are considered inappropriate or prohibited. This ensures that students and instructors are on the same page. This should be accompanied by outlining the consequences for failing to abide to the policy such as academic integrity violations.
  - **Guidelines for Using AI in Course Project Brainstorming**: Students may use AI to assist with brainstorming course project ideas, but the ideas have to ultimately come from students themselves. The AI generated ideas must be screenshotted or written out/cited if used to create your own idea.
  - **Regular Revision of AI Policies**: AI policies need to be flexible and not static. The course policies on AI should be reviewed and updated regularly to make it relevant to each class experience, expectation and intended outcomes.
- **Class 2 – PolicyCraft condition** (8 policies, excluding 3 without majority vote):
  - **Using AI to Combat the Language Barrier**: AI is acceptable to use for translation, improving grammar, and anything related to understanding language. However, AI translation that is directly translated from one language to another, specifically in reading responses or writing assignments,

shouldn't be submitted as one's own work. Translated words, phrases, and small sentences can be used though.
  - **Image Generation**: We can use GenAI to generate images to aid our essays or responses, with only restrictions being course guidelines (no explicit nature, etc). We should cite any use of GenAI for generating images.
  - **AI to be Used as a Helping Hand or Scaffold, not a Crutch**: Having AI capabilities to help provide insight rather than completely doing our work, specifically on helping hasten the process but not making a whole new design/idea from scratch. We suggest two conditions for how to avoid using AI as a "crutch". 1) If the task a student is getting support on is related to learning goals (i.e., it shouldn't be abstracted away), students should invest their own effort independent of a generative AI before they begin to use one. For example, as one might in an internship, students might spend 30 minutes on a task, and if they are still struggling, they might seek support from an instructor, peer, or generative model. 2) Students should engage with any ideas that a model produces or inspires in them. Rather than determining who owns an idea (of course still using attribution), course policy should focus on to what extent an idea offered a concrete learning experience for a student. If a student encounters a great idea while working with AI, and thinks critically about how they want to embark on a project based on that idea, that should suffice. Their ideas will continue to evolve. In contrast, simply copying and pasting idea without engaging with it (e.g., expanding upon it, considering how to actualize it, considering how one's perspective is related to it, etc., even if one does not edit the idea itself) would be less desirable.
  - **AI Usage Permitted for Coding Assignments**: Students may freely use AI for coding assignments with appropriate attribution. However, students should show understanding of any code AI has outputted.
  - **Developing a System that Uses AI**: Students may create systems that use AI, e.g., including a foundation model package or calling a foundation model API, so long as they make users aware of any potential harms and make efforts to avoid such harms in their design of the system (e.g., sharing sensitive chat data with OpenAI's API). Students should be aware of third-party APIs' data use policies in any cases in which they anticipate that users may share sensitive data.
  - **Usage of GenAI Tools should be Referenced and Cited**: In assignments or presentations, students should declare and cite the genAI tools and prompts that they have used to create content or help that they have received openly. While some cases involve grammar checks and simple paraphrasing for fluency, the original idea comes from the student. However, when students use GenAI to generate code or ideas for responses, it is essential to add a reference.
  - **Using AI for Rapid Prototyping**: With a general Idea of where to head, sometimes creating a lo-fi or hi-fi model can be hard for development due to a lack of skills expertise, no one's fault. But to bridge the gap and allow some fruition to catapult the project forward, AI can be used to help drive the initial run and allow ideas to get out of the base stages.
  - **Using AI for Citations**: Using Generative AI for creating citations is permitted. However, the student is responsible for checking the generated citation for accuracy and ensuring that all sources are properly cited in any assignment they submit.
- **Class 2 – baseline condition** (7 policies, excluding 12 without majority vote):
  - **Caution Against Misinformation**: Students are strongly encouraged to verify the accuracy and conduct fact-checks on any AI-generated content before including it in their assignments, as AI can sometimes generate false information, including fabricated quotes, citations, and research papers. Students should be responsible the accuracy of AI-generated information used in assignments.
  - **AI Use for Writing a Reading Response**: AI use in Reading Responses is limited to the following: (1) refining or clarifying your own written ideas, (2) assist with grammar, syntax, and minor rephrasing on an human written draft, and (3) cannot be used in generating original content or structuring arguments.
  - **Accountability of AI Responses**: When students use AI generated responses, they should be aware that the generated responses might not be 100% accurate. Students using information given by AI should be responsible for the accuracy of that information.
  - **Using AI for Group Organization**: AI can be used to assist in organizing group work such as drafting project outlines and summarizing meeting notes, provided that group members contribute to reviewing, editing, and finalizing these decisions
  - **Universal AI Attribution Policy**: If AI was used for a particular assignment, written notice must be given to the professor using the appropriate technology of submission for that assignment (e.g. comments in one's code if programming, the comment box of a canvas submission, etc.) outlining how AI was used in a particular work. This policy takes precedence over all other policy and is necessary to prevent any legal copyright/cheating issues.
  - **AI Usage Permitted for Coding Assignments**: Students may freely use AI for coding assignments, but must comment around (i.e. above and below) sections of code created using AI indicating AI attribution
  - **Cannot Create Segments/Personas/Archetypes**: GenAI can be used to assist in generating rough personas for general insights, pain points, or understanding the targeted audience. However, it should not be used to create detailed user segments or personas based on qualitative criteria like psychographics or behaviors, nor should it be the sole method for user profiling.

F Additional Detail on Data Analyses

F.1 Entropy Analysis

In addition to the mean entropy reported in Table 1, we fit a linear regression to analyze the impact of study condition (baseline = 0, PolicyCraft = 1) on policy entropy, controlling for between-class differences. As shown in Table 4 and Figure 11, entropy was lower for policies developed using PolicyCraft compared with the baseline ($p < 0.05$).

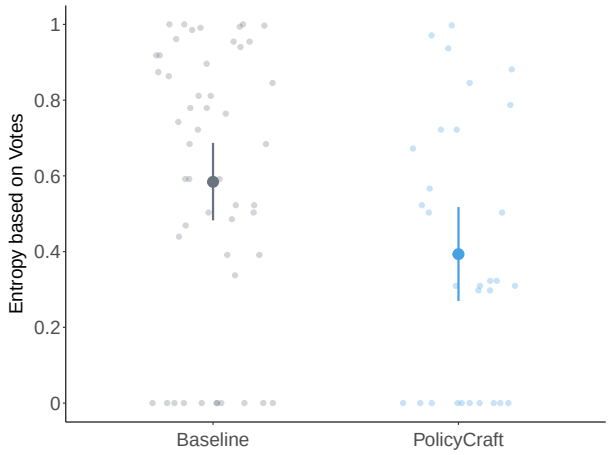


Figure 11: The distribution of entropy for individual policies calculated using the proportion of upvotes and downvotes each policy received. Lower entropy indicates a higher level of voting consensus.

F.2 Likert Data Analysis

In our analysis of participants' Likert scale ratings in Section 6.2, we employed both linear and ordinal regression to ensure the robustness of our results, considering the treatment of Likert scale data as either continuous or ordinal [74]. The appropriateness of using ordinal versus linear regression for the analysis of Likert scale data has been a subject of wide debate, with recent scholarship showing that each approach has complementary benefits and drawbacks [74]. In both of our models, Likert ratings are the dependent variable and study condition (baseline or PolicyCraft) is a binary independent variable. The class ID is included as an additional control variable, to account for potential between-class differences. As shown in Table 5, our findings are robust to the choice of ordinal or linear regression.

Table 4: Regression coefficients. Consensus was significantly higher (lower entropy) for policies developed in the PolicyCraft condition.

variable	estimate	std. error	t value	p value	
class	-0.1588	0.0815	-1.948	0.0550	
condition	-0.1930	0.0815	-2.368	**0.0204**	*

Table 5: Coefficients for both linear and ordinal regressions analyzing participants' ratings of the extent to which they could "easily understand why people agree or disagree with each other".

	variable	estimate	std. error	t value	p value	
linear regression	class	-0.2266	0.3276	-0.692	0.4916	
	condition	0.9066	0.3190	2.842	0.0060	**
ordinal regression	class	-0.1908	0.4575	-0.417	0.6768	
	condition	1.2513	0.4621	2.708	0.0068	**