ANALYZING AND DESIGNING FOR DELIBERATIVENESS IN ONLINE POLICY DISCUSSION

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Brian James McInnis

August 2019

ANALYZING AND DESIGNING FOR DELIBERATIVENESS IN ONLINE POLICY DISCUSSION

Brian James McInnis, Ph.D.

Cornell University 2019

Deliberation is a discussion-based practice that involves challenging a group of people to consider a policy issue by carefully weighing the diversity of perspectives held by members of the group. Deliberation theory expects that the practice contributes to an informed populous, well-reasoned decision-making, and mutual understanding among political adversaries. With such expectations, it is no wonder that deliberation theory and practices have been a source of inspiration when determining how to analyze an online policy discussion.

However, applying concepts from deliberation to analyze an online policy discussion is not trivial. There is a wide variety of discussion-based protocols that are referred to as deliberation, that involve varying degrees of facilitation, decision-making, and dialogue. As a result of this variety, there is no clear standard method for analyzing how people engage with topics, perspectives and others during a deliberation (referred to as the state of a deliberation, or its "deliberativeness"). Additionally some assumptions about the group membership and task of a deliberation are less relevant in an online policy discussion setting. As evidence of how these conditions complicate the practice of studying deliberativeness, the dissertation contributes a systematic review of the common analytic decisions for studying deliberativeness in online policy discussion.

A central argument in this dissertation is that the challenges involved with studying deliberativeness in online policy discussion can be addressed by

tightly integrating analysis with system design research. In an analysis it is important to consider how the design of a system might influence what participants are likely to contribute during an online policy discussion. In the first of two case studies, I present results from an experiment that demonstrate how key analytic concerns, such as the topic coherence of a discussion in the face of disagreement, are influenced by design decisions about how to order existing comments in a discussion. Analysis concepts might also offer a valuable source of inspiration for online policy discussion system design. In the second case study, I present results from an experiment that evaluates a process for introducing newcomers to an online policy discussion, that was inspired by analysis concepts, specifically meta-talk about conflict.

The residual benefit to the design "work" involved with operationalizing analysis concepts, like topic coherence and meta-talk, for an online context is that this work may help to narrow the gap between the theory and practice of deliberation and the study of deliberativeness in online policy discussion. In so doing, standards for operationalizing these concepts may emerge to ease the process of developing and evaluating theory about deliberativeness in online policy discussion. Insights about online policy discussion based on concepts from deliberation may also contribute back to the analysis and design of deliberation practices, whether in online, face-to-face, or even hybrid-online settings.

BIOGRAPHICAL SKETCH

Broadly speaking, the purpose of Brian's work is to help people collaboratively build insights around policy concerns. In practice, this means developing and evaluating public engagement technology through controlled experiments and in-the-wild studies, and then synthesizing what people convey during system design research into policy analysis and recommendations. While Brian's research involves modern crowdsourcing and online communities, he also draws inspiration from the theory and practices of deliberation.

Brian earned a dual Bachelors in Economics and History from the University of California at Davis. During college, Brian led Get Out The Vote campaigns and worked with state leaders to address education policy issues as a member of the California Preschool through Postsecondary Education (P-16) Council. Brian earned his Masters of Public Policy from Vanderbilt University's Peabody College of Education, where he studied issues related to the No Child Left Behind Act. Prior to joining Cornell, Brian worked at the RAND Corporation, where he studied a range of public policy issues—from the design of youth summer learning programs to predictive policing techniques.

Since 2013, Brian has been a PhD student in Information Science at Cornell University. In his graduate studies, Brian has continued to advance research into the ways that people build insights around policy concerns, by conducting a series of studies that involve Amazon Mechanical Turk (AMT) crowd workers (called "Turkers") in online discussions about issues related to the AMT participation agreement. This work with Turkers demonstrates a pattern of system design research coupled with policy analysis, that Brian intends to follow throughout his career.

To the Turkers at Amazon's Mechanical Turk.

ACKNOWLEDGEMENTS

There are several stories that I like to share when people ask about my committee. Early in my PhD program, Gilly Leshed helped to focus my energy by regularly holding up her hands and saying, “you were talking about this and now you’re talking about that. Choose one.” Gilly’s patience and thoughtful feedback at every step, has helped me to identify a program of research that I find truly rewarding.

Dan Cosley (aka. DanCo) has a knack for helping graduate students to evaluate their research ideas. When I have felt deep in the weeds with literature, Dan and I have taken strolls across the Cornell campus to synthesize and unpack the ideas, occasionally stopping for ice cream at the Dairy Bar. I look forward to doing the same with my own students.

In research and in life, it is easy to miss the forest for the trees. Since my first week at Cornell, Poppy McLeod has nudged me toward theory and concepts that help to situate my system design research in the broader context of communication and research about human behavior in groups. Beyond their academic lives, my committee members are athletes, dancers, chefs, musicians, and board game enthusiasts—friends who inspire me.

I am deeply grateful for my family. Throughout the ups and downs of dissertation writing, my Mom, Dad, Scott, Brooke, and Aunt Susan have helped me to weather the rough periods and to appreciate each milestone. I am also grateful to have joined Cornell with Jean Costa, Erica Ostermann, and Samir Passi. Our small, yet fierce cohort has been a great source of strength during these last six years.

Humbly, I realize that many people have contributed to the direction of my career. While there is not enough space in this dissertation to acknowledge ev-

ery other individual who has positively impacted my life, know that I regularly think about our time together in Ithaca, Washington D.C., San Diego, Chicago, Nashville, at the U.C. Davis Memorial Union, backstage at the Playhouse, close-hauled through the Bay fog, and so on.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

It is common in research to try to understand the social behaviors that emerge within a novel context by borrowing concepts from disciplines where there are established theories, practices, and analysis methods used to understand a similar context [199]. For example, online discussion systems offer a novel context for people to meet and talk with each other about public policy issues. Online systems are an exciting venue for policy discussion because of the potential to involve many people in discussions about broad policy concerns [34]. However, policy discussion forums also tend to be rude spaces that are often off-topic and lack meaningful engagement with disagreement [42, 210]. The social behaviors that emerge in online policy discussion are often compared against those that emerge in deliberation, which is "characterized by rational reasoning, postings on-topic, and reciprocity and respect toward other discussants" [8, pg. 38].

In order to study the social behaviors that emerge in a novel online policy discussion setting, researchers commonly borrow theory and analysis methods from the century-old practice of public deliberation. A deliberation is typically a brief and professionally moderated activity that involves challenging a small group of people to consider their diverse perspectives on a policy issue [78]. In order to analyze the quality of a deliberation process, deliberation scholars pay attention to the ways that participants engage with policy topics, perspectives, and each other over the course of a deliberation [108, 207]. Theory about deliberation expects that when deliberation is *done well* the process will yield well-reasoned decisions and will contribute to broad societal outcomes, such as an informed populous and mutual understanding among political adversaries [21, 25].

While deliberation and online policy discussion are similar in some ways, the assumptions underlying a deliberation are different than they are at most online policy discussions. For example, at a deliberation a professional moderator will follow the discussion and will help the participants to deeply engage with the policy topics by prompting conversation with a well-crafted question [160, 177]. By contrast, the large volume of commenting at some online policy discussions can make such active moderation a daunting task [61]. To account for such different assumptions—in group membership, facilitation procedures, and discussion setting—researchers must make some conceptual or operational adjustments when determining how to apply deliberation concepts to study an online policy discussion.

However, it is not merely a matter of mapping concepts from one context to another when trying to understand the social behaviors that emerge within a novel setting. Central to every academic discipline is an ongoing debate about the use and usefulness of theory, practices, and analysis methods associated with the discipline. Currently what constitutes an "established" method for analyzing a deliberation is still debatable [13, 21, 63, 107, 108]. This ambiguity about how to study the social behaviors that emerge during a deliberation has contributed to a theory-practice gap that has hindered empirical research about deliberation and the advancement of deliberation theory [170, 218]. Online policy discussion researchers grapple with this theory-practice gap when determining which version of a deliberation concept to use and how to operationalize the concept, whether for research or design in an online discussion context [156].

To investigate how online policy discussion research and design can more effectively use deliberation concepts the dissertation contributes a systematic literature review of the reported decisions that researchers make when doing so. As there is some debate about what are established deliberation analysis methods [13, 107], I chose to center the research around analysis concepts presented in Jennifer Stromer-Galley's [207] *Measuring Deliberation's Content* (MDC). MDC is a commonly used method that represents a broad set of deliberation approaches focused on rational argument and small group decision-making. The presentation of MDC also clearly maps theoretical concepts to operational decisions which is useful in analysis [199], and I also argue is useful in the design of discussion systems.

By focusing on just one analysis approach the systematic review sheds light on the variety of ways that a specific set of deliberation concepts have been applied and adapted by online policy discussion research. The review highlights how system design choices affect basic assumptions about policy discussion, in terms of the group membership, facilitation procedures, and discussion setting. While a deliberation analysis approach, such as MDC, assumes a policy discussion among a small group, that is moderated, and synchronously engaged with the discussion, there is a much wider variety of design possible in online policy discussion. To accommodate a wider range of design choices, new concepts and analysis methods have been coupled with MDC to account for different behaviors in online policy discussion.

To further investigate the interplay between analysis and system design in the practice of studying online policy discussion through the lens of deliberation concepts, the dissertation also contributes two case studies into this interplay. When analyzing an online policy discussion it is important to consider how the system design might influence what participants are likely to contribute to the discussion. In the first of two case studies, I present results from an experiment that centers on a common problem faced by online policy discussion as well as deliberation: Maintaining the topical focus of a policy discussion. Policy topics advance during a discussion as people reply and respond to each other along a common thread of subjects, but when the discussion is regularly off-topic or transitions from one topic to the next too quickly the discussion is unable to deeply engage with a policy issue [208].

However, maintaining topical coherence in an online policy discussion requires different considerations than deliberation. For example, a deliberation is typically restricted to a select group of people who are recruited because their experience related to a policy contributes to the diversity of perspectives on the policy held by the group [78, 108]. By contrast, many online discussions are open access and new participants are able to join at any point. While a newcomer may add a new perspective to the discussion, their new perspective often directs attention away from—rather than engage with—existing topics under discussion, thereby reducing the topic coherence of the discussion [33, 94].

One possible design lever to affect the topic coherence of newcomer contributions to an online policy discussion is curation: System designers can choose which comments are displayed, when, and in what order. In the case study, I examine how curating comments around particular positions (i.e., pro or con) on a contentious policy proposal might affect the likelihood that new contributions to the discussion will build on topics already in the discussion (called topic coherence). I found an asymmetric interaction between the curated position and a participant's own position, whereby participants with one specific position on the proposal were much less likely to add to the coherence of the discussion when it is curated with comments that oppose their position. The case study contributes system design recommendations related to the curation of topics where there are different/opposed perspectives in a discussion.

The first case study also builds on observations in the systematic review, by speaking to the challenge that researchers face when determining which version of a deliberation concept to operationalize for online policy discussion research. There are several deliberation concepts that are similar to the MDC definition of topic coherence. While a few articles in the systematic review apply the MDC concept of topic coherence directly [142, 189], the alternatives examine whether a comment is *relevant* [8, 94, 228] or *stays on topic* [204, 205] or is *related, but off-topic* [82]. Some definitions of what is *relevant* also impart value, as a comment that is less related to the current topic may be denounced as *irrelevant* or *off-topic* [178, 82].

Where the deliberative concepts of topic and topic coherence are widely applied through online policy discussion research, some other deliberative concepts are better known in theory than in practice; in the second case study, I present meta-talk as an example of such a concept. MDC describes meta-talk as talk about the state of the discussion that, "instead of advancing an opinion claim, is talk that expresses what the speaker thinks has happened or is happening and why it is happening in the discussion" [207, pg. 12]. During a deliberation, meta-talk is used to identify opportunities for consensus, conflict, or points that should be clarified. Statements like, "I think we all agree" or "I sense some disagreement around” help a group to transition from discussion about policy problems and toward vetting possible solutions.

In practice, an online deliberation moderator might coax a discussion forward with such remarks, but meta-talk is generally a rare phenomenon [207]. When an online policy discussion is regularly uncivil, occasionally a commenter will raise meta-talk to discuss another users lack of respect for others [14] or the general lack of civility at the forum [33]. However, in the systematic literature review I did not find the other forms of meta-talk described in MDC (i.e., consensus, conflict, clarification).

To investigate system design strategies that promote meta-talk in online policy discussion, I developed and evaluated a system inspired by *meta-talk about points of conflict* that I modeled after a protocol for teaching social studies teachers how to craft policy discussion prompts [177]. Borrowing concepts from crowd-writing systems [9, 134], I developed a process that structures the work to craft a discussion prompt into a brief onboarding task where newcomers to an online policy discussion synthesize and reframe two existing comments from the discussion. In the second case study, I present results from the evaluation, which demonstrate that writing and improving policy discussion prompts is a difficult task, but one that newcomers can perform reasonably well with appropriate structuring and clear guidance. In addition to demonstrating the value of using under-explored concepts in deliberation theory to inspire design, the case study also makes contributions to the design of crowd-writing systems.

A central argument in this dissertation is that the challenges involved with studying online policy discussion through the lens of deliberation concepts can be addressed by tightly integrating analysis with system design research. By *tightly integrating* analysis with system design, I refer to a cycle of online policy discussion research that applies analysis concepts from deliberation to examine the influence of a system design as well as research that uses system design to develop insights about how to operationalize these concepts in the novel context of online policy discussion. Through a cycle of testing and tinkering with concepts, like topic coherence and meta-talk, standards for operationalizing these concepts will emerge over time and may ease the process of developing online policy discussion practices and evaluating theory about deliberativeness.

In the process of pursuing research about deliberativeness in online policy discussion the work presented through the dissertation also raised several concepts and assumptions that are new to—or perhaps understudied in—deliberation research. In Chapter 7, I discuss the possible implications of the research for newcomer involvement in online policy discussion as well as the use of system design as a tool to explore assumptions about policy discussion, deliberative concepts, and their operationalization. I describe opportunities for future work that investigates how choices about the process of online policy discussion relate to deliberativeness and how deliberativeness relates to outcomes yielded by online policy discussion.

Online policy discussion might also offer deliberation theorists, practitioners, and analysts an opportunity to step back from the ongoing academic debate within their discipline. Black [10] describe *dialogic moments* during a deliberation as an important opportunity for people to let down the adversarial positioning commonly associated with debate by engaging in a process of mutual understanding through dialogue and narrative storytelling. The novel uses of deliberative concepts to analyze and design for online policy discussion may offer such an opportunity for dialogue about the theory, practices, and analysis methods that are commonly used to study deliberation. In this way, the "appropriation" of concepts that is common among academic disciplines might be reframed as an opportunity to reflect on the familiar and novel uses of a concept when applied to a new context.

The dissertation makes the following contributions:

1. *To online policy discussion system research*: (A) A review of how deliberation analysis concepts have been applied to online policy discussion research with recommendations about how the review can be used to guide online discussion research, (B) System design recommendations related to the curation of topics where there are different/opposed perspectives in a discussion.

2. *To policy discussion system research*: (A) Development of a process for crafting policy discussion prompts that could easily become part of a multi-staged deliberation process, where participants initially generate discussion prompts based on quotes from the last deliberation.

3. *To crowdsourcing research*: (A) Evidence of how personal perspective related to a controversial topic can influence the performance of crowd work related to the topic, (B) Further evidence of best practices around task clarity, (C) Recommendations about how to consider and respond to ethical concerns about crowd work.

4. *To deliberation research*: (A) Further operationalization of common constructs used in deliberation research for research in an online setting, including meta-talk and topic coherence, (B) recommendations based on online policy discussion research about how to address key questions for the

future of deliberation research and (C) analysis of places where existing deliberation theory research is relatively silent or does not fit well with online contexts.

CHAPTER 2

COMMON ANALYTIC DECISIONS FOR STUDYING DELIBERATIVENESS

2.1 Introduction

Deliberation is a practice of challenging a diverse group of people to examine their different perspectives of a shared problem [21, 76]. A lot of planning goes into organizing a deliberation. People are typically recruited to participate [67, 108] and before actually deliberating, organizers will brief the participants about the issues and will offer strategies for engaging with each other about the topic [78, 146]. A professional moderator will often monitor the discussion and add an occasional question [145, 160]. While there are many ways to organize a deliberation, participants often deliberate for a short period, as a small group, within a shared space, and in real-time [108].

The purpose of this chapter is to shed light on the decisions that researchers make when applying concepts from deliberation to study online policy discussion. In research about online policy discussion, the phrase "deliberation" is often synonymous with a high quality [70, 94, 159, 189], ideal [128, 2, 212], or democratic conversation [204, 205] that is "characterized by rational reasoning, postings on-topic, and reciprocity and respect toward other discussants" [8, pg. 38]. The promise of deliberation has inspired online discussion system designers to build novel tools that support decision-making and online community around a policy concern [34, 41] (see *ConsiderIt* [118], *Deliberatorium* [115], *e-Liberate* [195], *gIBIS* [36], *OpenDCN* [45], *RegulationRoom* [185]). In fact, there is so much activity in this design space that there have been numerous attempts to create a taxonomy for deliberation system design [1, 11, 34, 40, 43, 219, 179].

Despite this interest, there are few agreed upon approaches for evaluating the deliberativeness of an online policy discussion. This problem mirrors a well-known dilemma in deliberation research, where there are numerous discussion-based practices referred to as deliberation [10, 107, 108], yet few agreed upon methods for analyzing how participants engage with policy topics, perspectives, and others during a deliberation [13, 63], which has hindered theory development [170, 218]. This dilemma is recognized as a "theory-practice gap" in deliberation research.

This chapter presents a review of the common analytic decisions that researchers have made to bridge a related gap between deliberation and online policy discussion. By "analytic decision" I mean the choices that researchers make when determining how to study social behaviors that emerge within a novel context, such as online policy discussion, by borrowing concepts from a discipline like deliberation, where there are established theories, practices, and analysis methods used to understand a similar context. Analytic decision-making refers to the process of identifying and adapting a set of conceptual and operational definitions and assumptions to account for in an analysis [199].

As there are a variety of approaches used to analyze the communication that occurs during a deliberation [13, 107], the review centers on articles that reference Jennifer Stromer-Galley's [207] *Measuring Deliberation's Content* (MDC). MDC was selected to seed the literature review for several reasons. First, MDC includes an extensive coding scheme as well as a case study that offers a guide for how to perform this analysis. Second, MDC was designed to study deliberation among informed non-experts (i.e., laypersons motivated to talk about a specific policy issue). Online policy discussion is not limited to parliamentarians or elected officials [203], so we felt that the focus on informed non-experts would be generally relevant for researchers. Finally, the online policy discussion literature referencing MDC includes research about a wide variety of online systems and settings (see Table 2.1).

Type of System	*References*
Augmented Reality	[83]
Deliberation Systems	[62, 82, 86, 159, 169, 209]
Discussion Forums	[2, 8, 31, 85, 128, 142, 156, 165, 178, 205, 220, 228, 232]
Newspaper Websites	[15, 33, 69, 70, 136, 189, 204, 231]
Synchronous Chat	[211]
Social Platforms	*References*
Facebook	[94, 158, 189, 212, 231]
Twitter	[70, 125]
Wikipedia	[14, 227]
YouTube	[94]

Table 2.1: Articles that reference MDC [207] organized by type of system or social media platform to emphasize the breadth of online contexts within the MDC citation history.

Several themes emerged through the analysis, which are presented in the findings as four vignettes. A *vignette* is an evocative narrative, and in our case a short summary of relevant methods and results that offer context for the analytic decision-making. The vignettes examine how MDC concepts about deliberation have been operationalized (section 2.4.2) or adapted to study an online policy discussion (section 2.4.3). However, there are also reasons not to use MDC in an analysis, as there are missing concepts (section 2.4.4) and assumptions at play online that differ from those during a deliberation (section 2.4.5).

Each vignette concludes with a discussion about the lessons learned, which I then synthesize in the general discussion as common analytic decisions about

(1) dependent variables, (2) independent and control variables, and (3) the theoretical concepts that guide an analysis. The vignettes demonstrate that researchers have adopted deliberation concepts to analyze online policy discussion in a wide-variety of ways. While some deliberation scholars argue that such variety has hindered research in the deliberation field [170, 218], others argue that deliberation research demands multiple approaches to investigate deliberative concepts [13, 107]. I offer that the common analytic decisions might serve as an initial guide when preparing to study deliberativeness or when studying existing research about deliberativeness in online policy discussion.

The common analytic decisions developed through this chapter also provide a conceptual foundation for the central argument in this dissertation: *The challenges involved with studying deliberativeness in online policy discussion can be addressed by tightly integrating analysis with system design research.* In Chapter 3, I draw from the analytic decisions to introduce the Cornell e-Rulemaking Initiative (CeRI) RegulationRoom discourse architecture, which inspired the case study research presented later in the dissertation to demonstrate the interplay between analysis and designing for deliberative concepts (Chapters 5 and 6). As demonstrated through the dissertation, the process of design can lead to new insights about deliberative concepts and assumptions.

2.2 Definitions and assumptions for studying Deliberativeness

It is hard to pin down a definition for the term *Deliberation*. To quote Diana Mutz [170, pg. 525], "it may be fair to say that there are as many definitions of deliberation as there are theorists." Specifically, there is a wide range of norms for what a deliberation should be [21, 25] and many group discussion-based protocols that are all called deliberation (e.g., *AmericaSpeaks* [139, 138], *Community Literacy* [99], *Deliberative Polls* [68, 144], *National Issues Forums* [77, 101]). There are also many outcomes expected from deliberation, such as better decision-making [78, 108], public awareness [23], and mutual understanding [10, 63].

For that reason, the following terms are used through the chapter to separate the normative concepts of deliberation from deliberation practices and from the political science theory about how deliberation might be a central component of governance. However, a thorough discussion of each of the following terms is outside the bounds of the chapter.[1]

- **Deliberate**: *A normative concept*. To deeply consider a topic or to resolve a shared problem, by carefully weighting a diversity of perspectives [21].
- **Deliberation**: *A practice*. A process or activity intended to help people deliberate, by themselves or as part of a group [108].
- **Deliberative Democracy**: *A governance theory*. A model of governing where deliberation is a central component to decision-making [25, 78].
- **Deliberative(ness)**: *The state of a deliberation*. How people engage with topics, perspectives and others (if part of a group) during a deliberation [13, 106, 107].

This section presents a comparison of the analytic decisions related to six

[1] For more information about the norms and practice of deliberation, I highly recommend Karpowitz and Raphael [108]. Chambers [25] and Gastil [78] discuss the virtue, yet challenges to deliberative democracy as a governance theory. In back-to-back articles in the *Annual Review of Political Science*, Mutz [170] and Thompson [218] raise several concerns about the theory-practice gap in deliberation research. Black et al. [13] compiled a review of methods used to study deliberativeness in public deliberation, I draw comparisons among these methods through the Related Work.

approaches for analyzing deliberativeness (i.e., [57, 79, 95, 190, 203, 207]). The approaches were drawn from a comprehensive review of deliberation research methods, presented in Black et al. [13], that was intended to promote standards in the practice of deliberation research, by offering a "toolkit" of common deliberation measurement techniques. In the remainder of this section, I compare the assumptions as well as the conceptual and operational definitions that underlie each analysis approach.

2.2.1 Assumptions about deliberation

Group membership is restricted to promote diverse views of an issue

A common purpose for a deliberation is to craft a set of policy recommendations [108]. To generate these, participants are recruited and assigned to deliberate on a specific topic with a group of people and for a brief period of time [21, 144]. Membership is typically restricted to about ~15 people [57, 95, 190, 207], but in general, a deliberation group should be "a manageable size in which participants can see and hear one another" [21, pg. 400].

Recruitment often involves a randomized or stratified selection process [164] and an interest survey to identify a representative pool of participants with diverse perspectives [108]. Participants are typically selected because an issue effects their community (e.g., K-12 public school closures [57, 207]), but deliberation is also used to discuss broad policy concerns, like foreign policy [79, 95, 190]. The organizers of a deliberation will often prepare policy briefs and offer training to help people communicate with each other about the issues [99, 148, 164].

In several ways, group membership in a deliberation is similar to group membership in laboratory-based research [150]. The nominal groups organized for a laboratory-based study are often simple and generic systems, isolated from their embedding context, and with members who have no past or future together. Similarly, in a deliberation, group assignment procedures will often separate people with prior relationships and task each group to briefly engage with an abstract policy scenario [108].

Deliberation procedures play into the legitimacy of their outcomes

The period for deliberation is relatively brief, e.g., one weekend [57, 95], two 90-minute sessions [207], 1.5-2 hours [190], or 30-60 minutes [79]. A deliberation might also be part of a large event [108], coordinating hundreds of people (e.g., 466 [95], 623 [57]) and multiple deliberation groups (e.g., 57 groups [79], 30 [95], 23 [207], 5 [190]). While one group deliberates, several others might in parallel or will be preparing to deliberate in series [37, 138].

An important assumption about deliberation is that each group deliberates in isolation, independent of other groups [67, 144]. In practice, groups are often less isolated during a deliberation event. In fact, some deliberation procedures incorporate a series of small and large group activities [37, 138]. However, it is less clear how to analyze the policy conversations that participants carry on from the formal setting of a moderated deliberation and into the less formal lunch breaks or coffee sessions during a public engagement event.

Most of the analysis approaches indicate that moderators were part of the reported deliberation procedures (e.g., [57, 95, 190, 207]). However, few approaches also consider how the contributions of the moderator affect the deliberativeness of the policy discussion [207]. This is an important consideration, as moderator training is not well standardized [145, 160] and with even subtle cues, a moderator can bias the discussion toward an outcome [202].

A deliberation might yield recommendations, but the people recruited to deliberate are not often the same people who lead the recommended actions. Instead, deliberation outcomes are intended to inform the actions taken by policymakers, interest groups, or others affected by a policy concern. In this way, the outcomes from a deliberation are used by people who were not part of the conversations that culminated in support for the outcomes. For this reason, the perceived legitimacy of a deliberation process plays an influential role on the usefulness of its outcomes [108]. If people do not trust the legitimacy of a deliberation, whether how it was organized or facilitated, then the influence of any outcomes that result is mute.

The following sections discuss how the practice of deliberation is conceptually and operationally defined to study the deliberativeness of a deliberation activity.

2.2.2 Conceptual definitions

A conceptual definition describes a phenomenon with (a) enough information to distinguish the definition from other possible understandings of the phenomenon and (b) some possible indicators to determine if or when the phenomenon has occurred [199]. The following examples demonstrate how researchers use conceptual definitions to apply and engage with theory about the social behaviors that emerge during a deliberation.

Researchers converse with theory by defining concepts

Deliberative democracy scholars argue that in order to sustain a civil society the populous must remain informed about public matters by regularly taking part in rational-critical discussion about societal issues [21, 25]. Jürgen Habermas imagined that such rational-critical discussion would enter into every day conversation as people meet and informally challenge others' views. Such informal conversation can offer an opportunity for personal reflection through mutual understanding [88, 89].

Habermas called such community-level deliberation as enacting a *public sphere* and described several ideals for how community members should engage with each other through deliberation (called "Discourse Ethics"). The Discourse Ethics include rational critical argument, coherence and continuity of topics, as well as reciprocity, reflexivity and empathy as people actively listen and offer an equal opportunity to consider others' views. An analysis approach that closely maps to the ideals defined by the Discourse Ethics is the Discourse Quality Index (DQI) [203].

However, focusing the DQI solely on the Discourse Ethics introduced several limitations. Some norms are hard to measure, as an example "Habermas places considerable emphasis on the authenticity of claims, an aspect of discourse ethics that the DQI ignores completely because it is unobservable" [203, pg. 43]. Additionally, the DQI does not include humor in its conceptualization of deliberativeness, as "Habermas appears to view humor as a vice, [where other theorists] view it as a virtue, since it may contribute to openness in a debate" [203, pg. 43]. These examples illustrate the challenge of choosing what to exclude when translating theory into concepts.

An important tenet of Habermas's Discourse Ethics is that participants have an equal opportunity to engage with each other in rational reason-giving. However, people have different levels of ability to deliberate [109, 192]. Sanders [192] famously argued that the act of deliberation is undemocratic, and rather reinforces societal divisions that underlie many policy problems. When participants are not political equals, a discussion procedure that prioritizes equal time to speak may not effectively account for the rhetorical differences between powerful and dis-empowered [97, 192] or moderate and militant voices [196, 201].

In the analysis approach presented in Dutwin [57], Habermas's concept of equality is updated to include some of the concerns raised by Sanders [192]:

> *Equality*: "Equality requires not only equal access but equal discussions, deliberation where the discourse of certain individuals is lacking or otherwise suppressed cannot be deemed equal by any standard. Without equality, we cannot say that a public opinion formed by deliberation is truly by the public" [57, pg. 241].

The concerns raised by Sanders [192] are expressed in the way that Dutwin [57] qualifies equality as access to a discussion as well as equal power within the discussion and decision process. This example demonstrates how researchers use conceptual definitions to converse with the normative ideals expressed in the original text of theoretical works like the Discourse Ethics.

Deliberation in practice informs theory about deliberation

Rather than start an analysis from a normative ideal, like the Discourse Ethics, a few approaches inductively develop a conceptual understanding of deliberativeness by observing deliberation in practice [95, 190]. The role of narrative offers a useful example of how analytic and social group processes are at play during a deliberation [10, 190, 180].

In *Measuring Deliberations Content* [207] (MDC), the concept "Sourcing" is defined as the evidence that people use to support a claim, such as references to quotations from discussion materials (e.g., books, articles, government reports, speeches). The concept of Sourcing derives from the Discourse Ethics, which argue that rational-argument is not possible if the evidence people use to justify their claims cannot be reasonably verified [88, 203]. In an analysis that applies MDC, Sourcing is used to track whether participants support their claims with discussion materials or if they default back to sharing personal narratives—claims based on personal narrative are harder to contest during a deliberation [207].

However, narrative can serve purposes in deliberation that go beyond sourcing [12, 190, 180]. The act of telling stories that relate to a policy issue can foster dialogue during a deliberation [10, 180, 190]. Personal narrative can also offer space for people to "try on" alternate perspectives without the personal risk of taking an explicit stance [12, 180]. People also use stories to persuade others and to negotiate values in the deliberation [180]. The collaborative act of storytelling can also help groups to strengthen the social bonds among members [12]. For example, "collective storytelling" during a deliberation may enable a group to sustain a sense of moral community around a policy issue despite (or because of) their tense discussion [190].

To summarize, personal narrative is regarded as a type of evidence that people use to justify an argument, but in practice narrative can also add social value to a deliberation (e.g., fostering dialogue, perspective taking, collective storytelling) [12, 190] as well as narrative-argument (e.g., persuasion, negotiation) [180]. The multifaceted role that narrative can play in the practice of deliberation contributes to a broader conversation among deliberation scholars about whether rational-argument *should be* the dominant norm in deliberation theory [10, 63, 192].

A few analysis approaches address this question by including concepts that reflect rational-argument as well as concepts related to social group processes [79, 203, 207]. When applying an approach developed by Gastil et al. [79], for example, researchers are instructed to track the analytic rigor of the deliberation as one variable, and contributions to a group's social bonds and emotional atmosphere as another variable. When there are strong social relationships among group members, the participants may also limit what they say in order to avoid conflicts, and in doing so, move the deliberation toward a consensus that does not accurately reflect the diversity of perspectives among the group (referred to as a "false consensus") [105, 216].

There are many ways to observe how people communicate with each other during a group activity, like deliberation. The focus of the next section is how to specifically observe a deliberative concept by constructing operational definitions for an analysis.

2.2.3 Operational definitions

An operational definition provides a complete and explicit explanation of how each component of a conceptual definition will be measured [199]. In this section, I discuss common units (or levels) for analyzing deliberativeness and then discuss common decisions about how to operationalize the concepts of *Equality* and *Disagreement*.

Deliberative concepts are observed in specific units

Black et al. [13] describe deliberativeness at the macro and micro-levels of analysis. A macro approach considers the deliberation in total. For example, Gastil et al. [79] define a global measure of analytic rigor that involves expert trained coders studying the transcripts of several group discussions and then rating the analytic rigor of each group along several dimensions, e.g., "the group members identified a very broad range of solutions to the problem, the group carefully considered what each participant had to say" [79, pg. 30]. The expert responses are then aggregated to construct a scale score of analytic rigor.

A micro approach considers the deliberation in parts (e.g., thoughts, segments, threads). Some micro approaches suffer from a problem of how to analyze segments that include multiple speakers. For example, Hart and Jarvis [95] analyze the text in the 20 words of context before and after a plural first-person pronoun (e.g., "we", "us", "our"). Due to the automated nature of parsing each discussion into 41-word segments, some segments involve multiple speakers, some segments overlap, and despite the buffer of words around each token some segments still miss relevant context.

Other micro approaches record each thought [207], demand [203] or argument [57] uttered by each participant, so that each unit (ideally) encapsulates enough context to understand what a participant has meant by their statement. For example, in MDC [207] each *thought* is defined as an utterance (from a single sentence to multiple sentences) that expresses an idea about a topic, and shifting topics indicates a new thought.

It is interesting to note that different operational definitions for the same concept may not be correlated [79]. Each operational definition for a concept will reflect a distinct view onto the deliberation, with respect to the level of observation as well as the role that the observer plays in the deliberation (e.g., participant, moderator, researcher). For example, Gastil et al. [79] report no association between an expert coded measure of a group's analytic rigor and a participant reported measure of deliberation quality. While there is no reason why analytic rigor and deliberation quality *should* have a strong correlation, the fact that they are not well correlated raises some question about the qualities of a deliberation that different stakeholders find meaningful.

There are different approaches to operationalize a concept

As illustrated in the following operational definitions for *Equality* and *Disagreement*, even at the same level of analysis a deliberative concept might be operationalized in several ways. Equality in deliberation might simply mean that participants have an equal opportunity to be heard, but it might also consider their ability to influence the discussion or decision-making. A simple measure to compare the level of equality in a discussion is to track the number of thoughts each participant contributes [57]. This measure provides a way to evaluate equality as the rate of contribution by participant.

However, as a measure of equality, the rate of contribution can be misleading. For example, a low rate of contribution might indicate that a participant is more often silent, but not whether they were silenced by inequalities within the discussion. One approach is to pay attention to interruptions, as these patterns directly affect a speaker's ability to "participate freely in a debate" [203, pg. 27]. Another approach is to monitor the discussion for expressions of respect. Respect toward counterargument in particular, "is a necessary condition for the weighing of alternatives, which some view as an essential element of deliberation" [203, pg. 26]. The rate of contribution, interruption, and respect reflect different ways to operationalize equality.

Disagreement is another common deliberative concept that may be operationalized in several ways. For example, in the MDC [207] the concept of disagreement is operationalized to capture when a participant's *thought* contains a phrase, such as "I disagree, but" or poses a rhetorical question, which is an opinion in the guise of a question. This operational definition captures the "presence of disagreement," which reflects the MDC conceptual definition of disagreement as, "a sign that there is a problem in need of a solution a sign that there are participants in the dialogue with distinct views" [207, pg. 5].

As a contrasting example, researchers applying the DQI [203] examine the ways that participants engage with (show respect for) a counterargument. To do so, "engagement with counterargument" is operationalized as a four-level ordinal variable: (0) a speaker ignored a counterargument, (1) a speaker incorporates a counterargument in their speech, but degrades it, (2) incorporates neutrally, with no explicit negative or positive value, (3) incorporates and expresses positive value for the counterargument. Here the contrast between MDC and DQI is whether to observe *thoughts that state a disagreement* versus *how a participant responds when their thought receives disagreement*.

2.2.4 Common decisions about how to analyze a deliberation

Each approach offers a nuanced view into the practice of analyzing a deliberation with deliberative concepts (i.e., [57, 95, 79, 190, 203, 207]). Taken together, the review demonstrates that there are many assumptions and definitions to consider when determining how to analyze the communication during a deliberation [170, 218]. The value of multiple approaches to choose from, such as MDC [207] or DQI [203], is that it is reasonably straightforward to identify an approach that accommodates the assumptions and definitions for a specific deliberation context [13].

The process is less straightforward when applying deliberation concepts to a non-deliberation context. To illustrate the extent to which researchers struggle with bridging this conceptual gap, Medaglia and Yang [156] present research about disagreement in an online discussion forum. Before conducting the analysis about disagreement, the authors derived ten concepts for studying deliberativeness from a review of 11 articles about theory, 3 analysis approaches, and several literature reviews. The authors felt that this review was necessary because, “there is little agreement regarding how public deliberation might be measured empirically,” but ultimately adopted the definition of disagreement from the MDC [156, pg. 735].

However, it is not merely a matter of identifying the *right* deliberative concepts to apply. There are assumptions at play during a deliberation that may be less relevant in a non-deliberation context. For example, an online policy discussion is rarely facilitated and the audiences attracted to an online policy discussion do not often reflect the type of groups that are assembled for a deliberation. Furthermore, the design of an online system may afford forms of participant interaction that are less likely or hard to record during a face-to-face deliberation, which in turn may affect how a deliberative concept is adapted for a particular online discussion context.

As a way to think through the process of applying deliberative concepts to analyze a non-deliberation context, we call attention to the common analytic decisions among the analysis approaches presented in this section. Each approach provides details about the specific deliberation process associated with the approach, so as to clarify what assumptions about deliberation in practice might affect an analysis. Assumptions about the deliberation process are addressed in terms of the group membership, facilitation procedures, and discussion setting. While these factors affect the outcomes that a deliberation yields [13, 107], they also play into the perceived legitimacy of a deliberation process [108].

Each approach also offers a list of potential concepts to include in an analysis and several approaches discuss why specific concepts were excluded from the list. To address the gap between the theory and practice of deliberation [170, 218], a few approaches list concepts that reflect lessons learned from the practice of deliberation, such as narrative-argument [180, 190], in addition to concepts that draw from well-established ideals for deliberation [88, 89] (e.g., rational-argument, equality, disagreement).

Finally, each analysis approach offers instruction about how to observe a deliberative concept. When crafting an operational definition, approaches commonly address decisions about the level of analysis for observing a concept as well as whose perspective of the concept should be considered through the analysis (e.g., participant, moderator, researcher). There is also a fundamental decision about how a concept might be framed in the phrasing of its operational definition(s).

These decisions help to distinguish one deliberation analysis approach from another. Deliberation scholars can draw on these distinctions to identify an appropriate approach for a specific deliberation context. The central question that guides this chapter is: *What are the common decisions that researchers have made to analyze online policy discussion with deliberative concepts?* To identify the common decisions that distinguish analysis in each context, I chose to center this research around the online policy discussion literature that references one widely applied approach for studying communication during a deliberation: i.e., *Measuring Deliberation's Content* (MDC) [207].

2.3 Literature Review Method

2.3.1 References to *Measuring Deliberation's Content* (MDC)

As presented through the related work, *Measuring Deliberation's Content* (MDC) [207] is conceptually similar to other approaches that draw inspiration from the Discourse Ethics and argumentation (i.e., [73, 88, 194]). MDC defines deliberation as, "a process whereby groups of people, often ordinary citizens, engage in reasoned opinion expression on a social or political issue in an attempt to identify solutions to a common problem and to evaluate those solutions" [207, pg. 3].

The MDC definition of deliberation includes six concepts to capture how participants engage with topics, perspectives, and others over the course of a deliberation process, referred to as *deliberative elements*. The deliberative elements are intended to capture the ways that participants talk about and rationalize their policy concerns (i.e., Reasoned Opinion Expression, Sourcing), what topics and perspectives are raised during the deliberation (i.e., Topic, Equality), and how carefully the range of perspectives are weighed by group members (i.e., Disagreement, Engagement). The deliberative elements are defined by MDC, both conceptually and operationally through an extensive communication coding scheme.

To address the gap between the theory and practices of deliberation, MDC also incorporates an additional set of concepts that reflect the social and analytic intra-group processes at play during a deliberation, referred to as *forms of talk* (e.g., problem-talk, meta-talk, process-talk, social-talk). Problem-talk includes thoughts that express individual opinions, agreement or disagreement with prior speakers (or their positions), and talk that raises questions that advance discussion of the policy problems, perspectives, or proposed solutions. Meta-talk attempts to step back and observe what has happened, or is happening, and why it's happening, rather than advance an opinion or claim. Social-talk includes greetings, apologies, praise, as well as any other banter that might build (or harm) the social relationships among group members. Process-talk reflects the thoughts about the technology used for the deliberation and the deliberation process, such as what are the procedures and how to publicize the outcomes.

When applying MDC to an analysis, researchers are instructed to first partition the discussion content into units at the *thought* level of analysis, then categorize each thought as contributing to a form of talk, and then use the six deliberative elements to observe how social behaviors unfold through a sequence of thoughts. A common way to apply MDC in research is to calculate the total number of thoughts associated with each deliberative element and compare two or more deliberations by their distribution of an element. MDC includes a case study of the *Virtual Agora Project* as an example of how to interpret and report the deliberative elements.

I selected MDC to seed the literature review for several reasons. First, MDC includes an extensive coding scheme, with definitions for concepts that reflect deliberation ideals as well as deliberation in practice. Coupled with the coding scheme is a case study of the Virtual Agora Project, provided as a guide for performing the analysis. Second, MDC was designed to study deliberation among informed non-experts (i.e., laypersons motivated to talk about a specific policy issue). As an online policy discussion is often not limited to parliamentarians or elected officials, I felt that the focus on informed non-experts would be generally relevant for researchers. Finally, MDC has been referenced in research about a wide variety of online systems and settings (see Table 2.1).

2.3.2 Search and Selection Criteria

In July 2017, I worked with another researcher to conduct a citation history search for articles that reference MDC [207], using the Google Scholar and Pro-

Referencing MDC as an analytic approach				
Ref.	*Research*	*Users*	*Messages*	*Policy Issue*
[8]	Exp.	68	148	Parenting and adoption
[31]	Exp.	75	458	Education policy
[33]	Sample	1,073	6,444	Newspaper articles
[62]	Deploy	1,346	435	Degree conferral
[70]	Sample	NA	3,514	#climate, #dadt, and #dreamact
[82]	Deploy	450	451	Political reform
[94]	Sample	NA	7,230	White House communication
[142]	Exp.	54	476	Israel's security policy
[156]	Sample	4,735	10,990	National issues
[158]	Sample	NA	1,095	Political reform
[169]	Exp.	26	241	Gun control
[178]	Exp.	534	NA	Marijuana legalization
[189]	Sample	NA	1,000	Newspaper articles
[204]	Sample	212	300	Newspaper articles
[205]	Exp.	50	707	Parenting and adoption
[209]	Exp.	179	879	Education policy
[212]	Exp.	NA	2,403	Newspaper articles
[231]	Sample	25	1,580	Newspaper articles
[232]	Exp.	436	266	Textiles production

Table 2.2: Selected articles organized alphabetically and by whether the research applies MDC as a method or makes reference to the approach as related work. *Ref.* specifies the citation. *Research* indicates the type of study reported: i.e., a system deployment ("Deploy"), an experiment ("Exp."), a field study ("Field"), an analysis of a sample of the contributions to an online policy discussion ("Sample"). The number of *Users* and *Messages* included in a study (when specified) are recorded. *Policy Issue* offers a few words to describe the topic of discussion included in the study.

Quest search engines. The Google Scholar results returned 266 articles, and we found 64 through ProQuest, finding only 25 articles that were cross-listed (total 305 unique articles). Two researchers initially considered each article, by reading the title, abstract and introduction to determine whether the article might fit the following criteria:

- *Online discussion*: We excluded studies of face-to-face only discussion.

Analyzing online discussion with an alternative to MDC				
Ref.	*Research*	*Users*	*Messages*	*Policy Issue*
[2]	Field	387	NA	Political reform
[14]	Sample	70	282	Wikipedia policies
[15]	Sample	776	2,237	Daily newspaper articles
[69]	Sample	NA	42,179,238	Newspaper articles
[83]	Deploy	120	NA	Urban planning
[85]	Sample	25	1,699	Financial planning
[115]	Deploy [86]	160	5,003	Future of biofuels
[125]	Sample	100	388,875	National politics
[128]	Field	22	NA	Education policy
[136]	Field	3,470	NA	Editorial control in commenting
[159]	Deploy	107	1,212	Local waste treatment policy
[165]	Sample	606	1,211	Australia's 2010 federal election
[211]	Exp.	70	NA	Clean air Policy; Fourth Amend.
[220]	Exp.	557	NA	Carbon-dioxide storage policy
[227]	Sample	NA	2,655	Wikipedia policies
[228]	Sample	437	3,933	International issues

Table 2.3: Selected articles organized alphabetically and by whether the research applies MDC as a method or makes reference to the approach as related work. *Ref.* specifies the citation. *Research* indicates the type of study reported: i.e., a system deployment ("Deploy"), an experiment ("Exp."), a field study ("Field"), an analysis of a sample of the contributions to an online policy discussion ("Sample"). The number of *Users* and *Messages* included in a study (when specified) are recorded. *Policy Issue* offers a few words to describe the topic of discussion included in the study.

However, the selected articles do include studies that compare online versus face-to-face policy discussion [128, 159, 211], and test novel hybrid experiences [83].

- *Discussion activity*: We excluded articles that did not report information about participation in a discussion, such as the number of participants or messages. This step was to focus on articles that report an analysis of an online discussion, whether applying MDC or not.
- *Peer-reviewed*: We searched for articles that appear in peer-reviewed jour-

nals, conferences, or as book chapters. We excluded unpublished dissertations, technical reports, and conference workshop submissions.

These filters reduced the selected articles to 56, which were read in their entirety. With a more fine-grained review, we considered how each article applied MDC to the research and more critically evaluated the reported findings. The fine-grained review resulted in 35 selected articles (approximately 11% retained from the initial search). The 21 articles that were removed during the fine-grained review were primarily excluded because we found that they did not actually meet our selection criteria (i.e., online discussion, discussion activity, peer-reviewed) or that they did not involve a discussion about a policy, public, or political topic (a criteria that we only recognized during the fine-grained review).

The 35 articles reflect a breadth of methodological decisions. Tables 2.2-2.3 provide a high level overview of the selected articles, showing that 19 apply MDC to study an online policy discussion [207] (Table 2.2), while the other 16 reference MDC, in the background or conclusions, but apply an alternative approach (Table 2.3). A few articles apply MDC for analysis with only minimal modification [165, 205], but others adopt just a few of the concepts [31, 33, 94, 142, 158, 178, 209, 212, 231, 232], and several incorporate concepts from other analysis approaches [8, 62, 70, 156, 82, 189, 204].

As highlighted in the Introduction, the articles report findings based on a variety of online discussion systems and policy contexts (see Table 2.1). Sixteen of the articles are based on controlled experiments or system deployments, like the Virtual Agora Project. However, many are based on a sample of comments collected from an online discussion system operating in-the-wild.

2.3.3 Systematic Review Method

In order to identify analytic decisions from the selected online policy discussion literature, I worked closely with another researcher to conduct an open coding of how concepts have been applied and adapted by the research. For each paper we identified each concept, typically introduced in the background section, then noted the operational measurement, and methodological considerations to studying and reporting the concept, often described in the methods or findings sections.

For example, Blom et al. [15] argue that a discussion is not deliberative "if one or just a few people dominate the discussion" [15, pg. 2]. This conceptual definition reflects an assumption of equality in deliberation (see section 2.2.2). The following operational definition is used to distinguish frequent from less frequent contributors: "People that comment often (posts ≥ 7)" [15, pg. 4]. However, if frequent contributors often add civil posts then their dominance may not be detrimental to the discussion equality. The following operational definition is one way to distinguish civil from less civil posts: "[posts] contain[ing] content such as profanity, racial slurs, or shouting," add incivility to the discussion [15, pg. 6]. The authors conclude that frequent participants add more incivility than less frequent contributors.

Our review method recorded each concept (e.g., equality), noting the relevant operational definitions (e.g., a frequent contributor has contributed 7 or more posts) as well as whether and how the concept was reported by the findings. For example, by tracking common conceptual and operational definitions across the selected literature, we found eight articles that consider contributor frequency [14, 15, 33, 85, 94, 125, 159, 165] and 17 articles that evaluate incivility in a discussion.

In total, we identified 193 operationalized concepts from the 35 articles. This coding allowed us to identify common conceptual and operational definitions and assumptions when choosing how to study an online policy discussion with deliberative concepts (e.g., group membership, facilitation procedures, discussion setting).

Several themes emerged through the analysis, which are presented in four vignettes. The first vignette reviews how the MDC concept of *Reasoned Opinion Expression* has been applied with some modifications at the operational level (section 2.4.2). I then present research that adapts MDC at the conceptual level, in a vignette about *Disagreement* (section 2.4.3). Together, these initial vignettes capture common decisions that researchers make when determining how to apply MDC to analyze an online policy discussion.

However, I found that there are also reasons not to use MDC for an analysis. There may be concepts that are missing (section 2.4.4) or assumptions related to a deliberation that simply will not stand in an online policy discussion setting (section 2.4.5). I wrote vignettes III and IV to capture how researchers grapple with differences between deliberation and online policy discussion.

2.4 Findings

Before presenting each vignette (sections 2.4.2-2.4.5), I review the ways that researchers have used deliberativeness as a lens with which to examine online policy discussion. Specifically, I identified the following themes: (1) to weigh design decisions about a discussion process [142, 178, 220], (2) to evaluate outcomes that derive from online policy discussion [62, 82, 159], and (3) to consider how people discuss policy issues online when policy discussion is not the primary topic under discussion [85, 94, 228].

2.4.1 Common research objectives

A common use for deliberative concepts is to evaluate how the design of an online system affects discussion on the system. Wright and Street [225] coined the phrase *Discourse Architecture* [70, 178, 212, 231] to describe how features of a system, much like the architectural design of "parliament buildings, council chambers and the like, not to mention the electoral system which fills those spaces with representatives, affects the quality of the discussion and the nature of the debate" [225, pg. 853]. In examples from the review, a deliberative element, like reasoned opinion expression, is used as a dependent measure to evaluate the effect that design features have on discussion.

Discourse architecture studies often involve controlled experiments that expose participants to specific conditions, such as varying the discussion anonymity [8, 31] or synchronicity [205, 211], then present results about the extent to which exposure to a condition influenced the discussion. Quasi-experiments are also common. A quasi-experiment might compare the discussion at a discussion system before and again after a critical system design modification [69, 212]. When an experiment is not possible, discussion systems can be compared by analyzing how the discussion of a similar policy topic unfolds at each platform [70, 94, 189, 231]. In discourse architecture research, common analytic decisions include choosing what architectural element(s) to vary and which to control in order to draw comparisons.

Research can also leverage information about the user activity at a system to study how discourse architecture relates to deliberative elements, such as engagement and equality. Activity trace data typically include time stamped information about when, where, and how participants “click”, “hover over”, or otherwise use a system feature. For example, many commenting systems require participants to reply directly to a comment when they want to communicate with the comment author; the activity data related to this interaction can be used to construct a social network of the discussion. A social network analysis can reveal insight about the influence of specific actors [14] or about the cliques of people who communicate with each other [156].

A second purpose for analyzing deliberativeness is to evaluate the legitimacy of the outcomes that derive from an online policy discussion (e.g., [14, 62, 82, 83, 159, 227]). In deliberation research, public trust in the process of a deliberation is necessary to promote trust in the outcomes that a deliberation yields [108]. Much like deliberation research, I found examples of online policy discussion research that relates measures of deliberativeness to discussion outcomes, such as commitment to a decision [2, 62], willingness to reconsider a perspective [209], or familiarity with alternative perspectives [142, 178, 220] (called *argument repertoire* [23]).

The organizers of an online policy discussion often also care about how the participants perceived their discussion. For example, Escher et al. [62] report results from an online discussion about graduate degree conferral requirements. The report indicates that most comments were on topic and included a clear position, and “the debate contain[ed] no evidence of uncivil communication,” which was echoed in the post-discussion survey comments, as “a majority [of participants] judged the discussion as respectful and based on rational arguments,” [62, pg. 145].

Too much of a deliberative element might also raise questions about the discussion outcomes. For example, Monnoyer-Smith and Wojcik [159] report results from the online discussions during a public engagement that involved in-person events and multiple discussion systems. The analysis identified that the discussion at an online Question and Answer website was well-reasoned, but further investigation revealed that the high rate of reasoning was largely driven by a government agency that chose to moderate the site by issuing carefully worded fact-based answers to most of the user-generated questions.

A third use for deliberativeness is as an analytic lens with which to analyze sites not originally designed for policy discussion. This type of research typically presents a case study, by first identifying periods when policy topics are discussed and then describing the pattern of discussion, whether it was deliberative or if other discourse norms were at play (see [85, 128, 158]). For example, policy discussions emerge in social media. However, the *publicness* of social media by comparison to less public settings, like anonymously commenting on an article at a newspaper website, “may discourage users from openly expressing their views and opinions, thereby reducing the quality of deliberation that might occur” [189, pg. 547].

Another focus for this type of research is policy discussion that emerges in non-political discussion settings. Ray Oldenburg [173] developed a theory about physical locations, like the café, a beauty parlor, or the local tavern where people gather informally and voluntarily engage in discussion about current events, while providing each other the opportunity to question, challenge and form opinions about a policy, individually and collectively. Oldenburg referred to these locations as *third places*, which offer a different type of policy discussion than is often possible at home with family and friends or at work (referred to as first and second places).

Scott Wright [226] adapted the concept of third place to describe online discussion settings as *third spaces* [84, 226], where policy-talk can and does emerge amid talk about non-political topics, such as financial planning [85], cricket [228], reality television [84], and parenting [128]. A case study about a third space will often describe the ways that people talk about policy issues within a specific discussion setting, such as their use of metaphor [128], informal moderation [85], and any non-deliberative norms also in the discussion [70, 228]. By examining policy discussions that emerge organically, we gain a greater understanding about how discourse architecture and social activity at an online system contribute to policy discussion and to its related outcomes.

With these common reasons for studying deliberativeness in mind, the following four vignettes review common analytic decisions when determining how to apply and adapt deliberative concepts to analyze an online policy discussion.

2.4.2 Vignette I: Applying deliberative concepts to an analysis

Online policy discussion research that has applied MDC [207] has adopted only some of the conceptual definitions (called Deliberative Elements). The six deliberative elements defined by MDC to measure the content of a deliberation include (see section 2.3.1): Reasoned opinion expression, Sourcing, Disagreement, Engagement, Equality, Topic. The deliberative elements are typically dependent measures, used to determine how factors related to the design of an online system or to the activity at a system affect a specific deliberative element.

The first vignette examines how *Reasoned opinion expression*, in particular, has been widely adopted in the analysis of online policy discussion.

Reasoned opinion expression

Twelve of the articles applying MDC as part of the research methods include reasoned opinion expression in the analysis [8, 31, 33, 62, 82, 94, 142, 178, 189, 204, 209, 212]. For an opinion expression to be reasoned, the expression must include a source of evidence to support the opinion, "that can be observably confirmed or empirically denied or appeals to a shared normative ground" [207, pg. 4].

Our analysis found that *opinion expression* is routinely operationalized as a simple dichotomy between broad political belief systems (e.g., conservative versus liberal [189] or progressive [70, 165]) or as a stance on an issue (e.g., support, neutral, opposition) [142, 156, 158, 178]. However, the purpose for recording an opinion is most often to study disagreement, rather than to specifically evaluate whether the distribution of opinion expression during a discussion was balanced or biased toward one side or another [189].

Operationalizations of the *sourcing* to support an opinion are fairly consistent, tending to treat it as either a binary [33, 94, 142, 178] or as a categorical variable [82, 204, 212]. Interestingly, in either case the majority of comments to an online policy discussion often do not include justification. For example, in a comparison of the comments to the White House page on Facebook and YouTube, 64.9% of the posts to Facebook included a claim, but did not support the claim with justification, while the lack of justification was significantly higher on YouTube (71.1%) [94].

Our analysis found that several studies examine the effects that similar independent variables, such as discussion topic and platform, have on reasoned opinion expression; however, their operationalization varies. For example, a discussion topic can be classified as *controversial* [8], *hard* [70], *sensitive* [94], *serious* [33], or by a *damage factor* [231, 232]. Inconsistency in the operationalization of similar concepts can complicate the process of drawing inferences from existing research. For example, a controversial topic does influence reasoned opinion expression, but it is less clear how, as we find evidence that controversial topics yield fewer reasoned claims than less controversial topics [8], yet also find evidence to the contrary [94].

Another common independent factor is the discussion platform, but some quasi-experimental controls are necessary to compare differences in reasoned opinion expression at platforms in-the-wild (e.g., Twitter [70], Facebook [189]). Common factors to control for include the discussion *topic* and the *state of an issue*. To control for the discussion topic the research might center on all comments in response to a specific article [189], a specific policy [70], or on the mass communications from a specific person or entity, such as the "White House social media presence" [94].

However, opinions about an issue and the range of sources available will shift as the issue continues to evolve. To account for the state of an issue, the discussion data should be collected to reflect a set period in the issues history (e.g., a few weeks [189], a few months [70, 94]).

Vignette Discussion

The vignette speaks to two types of analytic decision-making when determining how to apply an existing deliberative concept to an online policy discussion. The first is about how to operationalize a deliberative concept for research and the second is about how to examine or control for effects related to other variables, like topic or platform.

Not unlike other deliberative concepts, reasoned opinion expression is observed as the combination of sub-concepts: i.e., opinion expression, sourcing. We find some variety in the ways that each sub-concept is operationalized, for example in some research "neutral" is a valid opinion [62, 82, 156], but elsewhere the opinions that are not on one side or another are "unknown" [70] or "unclear" [189]. Being able to identify a left-right stance is fundamental for studying disagreement, but the reasons or sources that people with a less identifiable stance bring to a policy discussion might be quite different than those added by people who have a definite opinion.

At the same time, people express opinion in the sources that they reference. Consider how sourcing relates to reasoning in MDC: Conversation was considered informed when participants "used a shared resource to support claims," such as the background materials prepared for the deliberation [207, pg. 11]. When people deliberate they use reasoning to affect how other people interpret a shared artifact (e.g., a policy, a book, a video), which is similar to how people use narrative to consider alternate perspectives of a shared experience [190]. By contrast, claims that are supported by an external source are hard to contest until other members of the discussion are familiar enough with the source to apply it with their own reasoning [126]. For this reason, using a binary variable to capture whether a claim includes a source or not may miss whether referencing a source has contributed to an informed discussion or not.

Several factors, such as the discussion topic, platform, and the state of an issue, can affect the availability of sources as well as what opinions people are willing to express about an issue. Our analysis identified some rather basic inconsistencies in the terms used by several studies to characterize a discussion topic. For example, it is less clear what makes a topic more controversial [8] than it is serious [33] or sensitive [94]. There is likely some value in the creation of bespoke concepts for research into a novel context, but as research about the context accumulates—as it has in public deliberation—it is likely valuable to pause and synthesize similar concepts into standards that guide future analysis [13].

In this review of reasoned opinion expression, there are some practices that might be considered standard, such as the steps to control for the topic and state of an issue when comparing policy discussions at different platforms.

2.4.3 Vignette II: Adapting deliberative concepts

MDC defines disagreement as, "a sign that there is a problem in need of a solution, a conflict in need of consideration and resolution [and] that there are participants in the dialogue with distinct views on a particular issue" [207, pg. 5]. Note that this definition characterizes disagreement as an opinion expression that, when expressed, offers certain benefits to the discussion. MDC offers the following postulates about the presence of disagreement [207, pg. 5]:

- The presence of disagreement indicates that the discussion participants are not homogeneous in their viewpoints [207] (i.e., a diversity of perspectives)
- If the discussion participants are not homogeneous in their viewpoints, then the discussion is less likely to polarize [215]
- When exposed to views that are different than their own, people are more likely to have their own views strengthened in a more rational way [23]

While some of the articles that examine disagreement apply the MDC definition, most do not. Instead, most of the articles that incorporate a measure of disagreement conceptualize disagreement as a form of participant engagement, which implies different operational definitions, units of analysis, and theoretical expectations than conceptualizing disagreement in the way that MDC recommends, as a form of reasoned opinion expression.

Disagreement as a form of Reasoned Opinion Expression

A total 49 (~25%) of the concepts identified by the review relate to disagreement and several apply the MDC definition. Without access to survey data, voting patterns, or other forms of disclosure, researchers rely on what participants convey in their commenting to infer what stance a participant has taken on an issue. A common approach to infer a participant's stance is to calculate their average stance based on a review of all their prior statements about the issue (e.g., support, neutral, opposition) [70, 82, 156, 189]. This approach has limitations as it implies that a participant's expression of agreement or disagreement with an issue will, for the most part, be stable throughout a discussion.

Expressions of disagreement during a discussion can also be aggregated to draw inferences about the discussion. For example, the *rate of disagreement* is a common indicator that a diversity of perspectives have been expressed during a discussion [62, 82, 158, 207, 209]. As an example, Molaei [158] discuss how norms of politeness can yield polite disagreement online, presenting a case study about the diversity of perspectives in a series of discussions at a Facebook group focused on politics in Indonesia. More than half of the Facebook posts expressed agreement, yet only 15.5% (N=170) expressed an opposing view. The research concludes that, "[t]he low rate of disagreement with the main idea of discussion indicated some polarization of opinions in the forum" [158, pg. 497].

The analysis found that it is less clear how to interpret a rate of disagreement in terms of perspective diversity or polarization. Molaei [158] and Stromer-Galley [207] arrive at different conclusions about what the disagreement rate indicates. Molaei [158] characterize a discussion with 15.5% disagreement as showing "some polarization of opinions," though Stromer-Galley [207] finds 5.6% (N=351) disagreement to indicate "some heterogeneity in the views expressed, and that participants were hearing divergent perspectives," limiting the potential for polarization [207, pg. 18]. This highlights a Goldilocks-*like* dilemma when evaluating conceptual relationships [17]: How much disagreement is too much, too little, or *just* right to fend off polarization?

Researchers have conducted reflective interviews [231] and experiments to examine how people react when exposed to different doses of disagreement [178, 232]. For example, in a re-analysis of the Virtual Agora Project, it was reported that when exposed to either high-low or low-high combinations of agreement and disagreement during the online discussions, participants were more likely to be satisfied and willing to participate in future deliberations than when exposed to a balance of agreement and disagreement [209]. In other words, some people actively avoid disagreement, but others come looking for it.

Disagreement as a form of Engagement

Most of the articles in this survey are not about the presence of disagreement, but instead focus on the participant engagement related to disagreement. For example, Medaglia and Yang [156] examine whether participants in an online policy discussion tend to talk more with people who have a similar perspective or with people who oppose a perspective. The research involved constructing a social network from each reply and response during the discussion, then studying how the social network evolved as the discussion progressed (see also [14]). Conceptualizing disagreement as a form of participant engagement can reveal factors that may prevent opportunities for disagreement [156], such as participant attitude, a controversial topic, or structural aspects of the social network.

When people do join in a disagreement, it is important to consider the manner with which they do so. For example, the following describe the reply-and-response units of analysis for studying whether and how people acknowledge an expression of disagreement [204]:

- **Opinion expression**: Participant *A* makes an initial statement A^1
- **Expression of disagreement**: Participant *B* responds to A^1 with disagreement B^1
- **Engagement with disagreement**: Does *A* acknowledge B^1 and if so, how?

When people do acknowledge a disagreement, it is often with hostility [204, 211]. A variety of terms are used to characterize the tone of a disagreement: With *soft* or *bold* words [82, 211], *politeness* [158, 204], or by the *type of question* [70, 142]. In face-to-face situations, people can express "more bold disagreements while using non-verbal cues to soften those disagreements," such as eye contact or smiling [211, pg. 16], but in text-based communication people may rely on profanity to convey the valence of a position [69, 70, 211].

The design of an online system also influences the way that people are exposed to disagreement. For example, the commenting systems at a newspaper website typically allow for a threaded response, meaning that people can reply directly to a specific comment, which can promote a topically coherent engagement with disagreement [204]; whereas, in synchronous chatting, disagreement can be hard to coordinate and sustain because of the presence of multiple topics and the tendency of people to avoid contentious topics [211]. These aspects of the system design factor into decisions about how the concept of disagreement is operationalized as a participant engagement, e.g., a threaded reply-and-response [204], unthreaded discussion [211], discussion curated by distributed moderation [33], discussion within a hashtag [70].

Vignette Discussion

Our analysis found that most of the articles that examine disagreement do so by considering how people *reply-and-respond* to each other, rather than what *thoughts* people share about their stance on an issue. This shift in the units for analyzing disagreement implies that most research has appropriated the concept of disagreement in a way that has led to an operational definition that is quite far from MDC, which may mean different strategies to interpret and theorize about disagreement. In this vignette, I discuss the analytic decisions related to how observations of a deliberative concept might be interpreted.

Numerous social phenomena are observed in units of expressed disagreement. Expressions of disagreement can be aggregated from a thought-level to a discussion-level, so as to examine perspective diversity during an online discussion [207]. Similarly, participant interaction with disagreement can be aggregated to investigate reply-and-response patterns of participant interaction as a social network [14, 156]. A micro-level lens might also be applied to examine the tone or politeness associated with acknowledgement of disagreement [204]. The simple phrase "I disagree" might add value to the analysis of several social phenomena.

The analysis found that several factors complicate the practice of interpreting observations of disagreement. Researchers often have limited access to the people in an online policy discussion. During a deliberation, it is common to poll participants so as to analyze individual differences in stance and how positions shift during a deliberation [67, 138, 144]. Identifying a participant's stance by what they communicate does have limitations, while periodically polling people about their position may itself have an impact on the course of a discussion [105].

I also note that researchers may interpret thresholds of a measure differently, as Molaei [158] and Stromer-Galley [207] arrive at different interpretations of what a rate of expressed disagreement indicates about polarization. It is not hard to see how a rate of expressed disagreement might be interpreted in various ways. Several scenarios might yield a high rate of expressed disagreement: e.g., an equal amount of disagreement for each topic, or a deep level of disagreement with a specific topic, or the presence of a single high-frequency dissenter. However, each scenario implies a different distribution of perspectives and positions in the discussion. Disaggregating measures like the amount of expressed disagreement by topic, perspective, or period better reflects, what I see as, emerging standards about how to compare discussion at separate platforms by controlling for the topic and state of an issue (see section 2.4.2).

Although I do take issue with applying a disagreement rate as a substitute for missing information about how personal positions may shift in conversation. Group polarization refers to the tendency of a group to adopt positions that are more extreme than those initially held by the group members [215]. Two social mechanisms help to explain this tendency: (1) people receive social confirmation of their views when they hear other people echo those views, and (2) a view with few counter-arguments is perceived as more persuasive than a view with many. In light of this, a rate of expressed disagreement does not adequately reflect whether, through discussion, a diversity of perspectives have been introduced or if the group members have coalesced around a position that is more extreme than what each person might agree to otherwise.

It may be that some analyses that are common in deliberation, like examining group polarization [215], are tricky to implement in an online discussion context. Online systems are designed to expose participants to disagreement in ways that are different than in deliberation. Common strategies to affect disagreement during a deliberation, like smiling or eye-contact, are less available online [211]. Online systems may also introduce features intended to ease disagreement [168] or to communicate alternate views of an issue [83, 86], in ways that are less possible during a deliberation. While the design of an online system may affect how observations of disagreement are interpreted, the people who participate in online policy discussion may also value being exposed to disagreement differently than the people who participate in deliberation [129, 166, 209].

2.4.4 Vignette III: Adding new deliberative concepts

Research that examines *Incivility* reflects a pronounced deviation from MDC. Where earlier vignette's present research that has applied or adapted deliberative concepts—sections 2.4.2 and 2.4.3 respectively—the focus of this vignette is on ways to observe discussion behaviors that are not defined by MDC.

Incivility

In the review, 40 concepts (~21%) relate to incivility. Incivility is a focus of online policy discussion research, yet MDC [207] does not mention it. In fact, a quick word search will confirm that the terms "civility", "tone", "respect" and "politeness" do not appear anywhere in the article, though civility is a core component of the deliberative democracy theories that MDC references (e.g., [74, 88, 194]). MDC does define social-talk as talk that includes those "informal, sociable interactions that help groups to cohere," but the types of disruptive talk that erode social cohesion may have been less common in the deliberation data that was used to develop the analysis approach [207, pg. 31]. I raise this point as an acknowledgment that designing an approach for analysis inherently means making choices about what concepts to exclude.

To evaluate the civility of a comment, a researcher will consider the target of the message and it's valence in tone or in politeness. For example, cursing to emphasize a reasoned opinion about an issue might not be disrespectful [94, 211], but when people use foul language to disqualify others and their arguments, the interaction is disrespectful [82, 211] and might be classified as a personal attack [15, 33]. Civility is a tricky concept to operationalize because it involves making a value judgment about whether an expression that includes harsh language has the intent or effect of failing to show respect for another person or group in the discussion [211].

An alternative approach to operationalize the concept of civility is to pay attention to the work that moderators perform to manage the level of respect in an online policy discussion [62]. It is worth noting that MDC does include an approach to analyze how moderator contributions influence a deliberation, but we found no example of its use. However, the work of moderation can be less visible at times. For example, Strandberg and Berg [204] report that the discussion at an online newspaper website was mostly civil (98.7%), but suggest that this finding may be inflated due to the "post-moderation used by the [newspaper] which might be dis-proportionally reducing impolite and rude comments" [204, pg. 137].

While the moderation practices at an online policy discussion may offer some insights about the level of civility, the mere presence of a moderator may also influence the level of civility. Stroud et al. [212] present an in-the-wild experiment, varying whether the discussion at the Facebook page for a local television station was moderated and if by a publicly recognizable reporter or a generic newsroom representative. The study argues that the presence of a reporter in the discussion increased the salience of participant identity as part of the *reporter's audience*, which contributed to a more civil discussion than when a generic representative of the newsroom participated or when the discussion was not moderated.

The discussion may also respond to an uncivil comment. Occasionally people talk about the level of incivility in their discussion, which researchers observe as a type of meta-talk (see [14, 33, 204]). Recall that MDC defines meta-talk as talk that attempts to step back and observe what has happened, or is happening, and why it's happening, rather than advance an opinion or claim. For instance, Coe et al. [33] found that meta-talk was rare at an online newspapers website, but the few instances all expressed frustration with the level of incivility in the comments.

Many commenting systems also include *up/down* voting as a way to curate the discussion by distributed moderation [123, 124]. System features, like voting on comments, enable people to silently signal their (dis)approval of uncivil comments [33]. While voting on a comment is not meta-talk, the transparency of comment voting patterns at an online policy discussion might elicit meta-talk about the discourse norms prioritized by the discussion system.

Vignette Discussion

Incivility and Disagreement (see section 2.4.3) share some conceptual components. Both incivility and disagreement include a *target* and a *tone*, though what makes civility tricky to operationalize is the level of *respect* expressed by those components [211]. Rather than define a single operational term for respect, researchers might code for related concepts. For example, Stromer-Galley et al. [211] independently record the *boldness* and *politeness* of disagreements. This analytic decision to triangulate several indicators as a proxy for *respectfulness* for example, might be useful when working with other hard to operationalize concepts.

Civility can also be operationalized in terms of the actions that participants take in response to an uncivil comment. This is similar to how *engagement with disagreement* is operationalized (see section 2.4.3), in that the research focus is on the ways that participants respond, such as by raising meta-talk [14, 33, 204] and with moderation [62, 211], rather than focus on the uncivil content alone. This operationalization also highlights the natural delay between the point at which an uncivil comment is posted and others in the discussion respond. As a result of this delay, some common responses to incivility, such as comment deletion [48], might remove the content without also resolving its impact on the discussion.

Discussion system design decisions can make the response to an uncivil comment more and less visible. Stuart et al. [213] present a framework for thinking about *social transparency* in online systems, which is defined as "the availability of social meta-data surrounding information exchange" [213, pg. 451]. For example, incivility in online policy discussion is often related to the “identity transparency” (or level of anonymity) available at a system [8, 69, 94]. The vignette highlights other types of social transparency as well, such as the transparency of participant interactions with a system (e.g., up/down comment voting [33]). The content of an online discussion can also be altered without notice, as in a moderator deleting a comment [204], which relates to the content transparency of social activity.

Moderator activity is highly visible during a face-to-face deliberation. Moderators help the participants to know each other, to feel comfortable sharing, and to resolve differences productively, such that the group is able to address many of their deliberation prompts within the time allotted [145, 160]. With few exceptions [61, 86, 212], such human moderation is rare in online policy discussion. However, system design can make the social and policy analysis activities of participants and groups during an online policy discussion more and less visible.

2.4.5 Vignette IV: Introducing new assumptions

In practice, deliberation is a brief opportunity to consider a diversity of perspectives on a policy issue. Deliberation moderators prioritize equality of participation and perspective, so that the group has time to weigh a range of perspectives within the brief period allotted to address the deliberation objectives [145, 160]. *Equality of access* implies that participants have an equal opportunity to participate. Similarly, *equality of perspective* implies that each perspective has an equal opportunity to be heard and considered during a deliberation [21, 109, 207].

These ideals for equality are rarely met in online policy discussion. While a deliberation is typically brief, a policy issue may remain in public focus indefinitely. As the time available for discussion increases there is more opportunity for people to develop identity [2], roles [85, 165], and collective actions related to their policy discussion [128]. In this vignette, I discuss how equality of participation and perspective in online policy discussion are affected by the social behaviors of high frequency contributors.

High Frequency Contributions to Equality

Several articles in the review compare the contributions and roles enacted by high and low frequency participants (i.e., [14, 15, 33, 85, 94, 125, 159, 165]). These articles present examples of how the contributions of highly active participants, in particular, affect the equality of access and perspectives at an online policy discussion.

One way to evaluate equality of access is by calculating the average rate of participation (see section 2.2.2). Most online discussion communities exhibit a long-tail distribution of participation, which indicates that most of the participants participate briefly, yet most of the discussion is among a numerical minority of people who participate frequently [153]. A discussion with a low, even one-time, average rate of participation is considered more egalitarian than a discussion with a participation distribution that is skewed toward frequent participants [94].

Another way to evaluate equality of access to a discussion is by the presence of incivility, as a regularly disrespectful discussion can feel less inviting (see section 2.4.4). Both low and high frequency participants are sometimes to blame for incivility, but there are different explanations for what motivates their uncivil behavior. A common explanation for why low frequency participants contribute comments that are uncivil is that they are "drawn to comment by a particularly upsetting article that encourages an uncommon-and uncivil-post" [33, pg. 673]. In other words, contextual characteristics associated with an article elicit incivility from infrequent participants, such as a controversial topic, author, or source [28, 33].

By contrast, high frequency participants may habitually stoke an uncivil tone, as they "may, under some circumstances, be actively cultivating conflict" [14, pg. 615]. Due to their volume of participation, a frequently uncivil participant can have a detrimental impact on the discursive norms at an online policy discussion [15]. While a system designer might address a rash of uncivil newcomers by forgiving some early infractions [28], the prevailing recommendation about how to address incivility among frequent participants, based on this review, is to reduce their influence by recruiting new people to the discussion [14, 15].

Due to their volume of activity, high frequency contributors may also drown out all but their own perspective—contributing to an inequality of perspectives. Larsson [125] analyzes the political positions of the 100 most frequently *re-tweeted* contributors (referred to as elite users) in political hashtags on Twitter. While each hashtag provides a venue for citizens to engage with policy issues, "a significant amount of activity is undertaken by established political actors as well as niche political groups," [125, pg. 56]. Where a political actor will promote their perspective by contributing frequently, political advocacy groups will promote a perspective by coordinating a message among their members, who may themselves be infrequent contributors. The dominant role played by political actors and groups can marginalize the perspectives of non-elite users [125, 159].

However, the contributions of high frequency participants are not always detrimental to equality. For example, Graham and Wright [85] study the contribution patterns of super-participants at a personal finance forum, finding that most of their contributions are rational-critical replies and that super-participants will synthesize and clarify arguments, adopting a *de facto* discussion moderator role. Similar behavior was reported at a collaboratively written political blog, where active members add to policy discussions by, "looking for and providing more detailed analyses when needed," thereby enhancing the shared understanding of a policy issue, and supporting the blog authors as "co-experts" [165, pg. 35]. In both examples, frequent participants contribute in ways that promote equality of access and perspective.

Vignette Discussion

The frequency of participation and perspective during a deliberation is managed by a facilitation procedure (and often by a moderator) [145, 160]. Additionally, the brief period of time available for deliberation limits the potential for participants to participate with high frequency. These assumptions about group membership are different in online policy discussion, where membership evolves as participants enter, exit, return, and are remembered by the discussion. For this reason, it is important to consider the ways that people join [159], identify [2, 125], and become socially active within an online policy discussion [85, 165].

Due to their volume of activity, high frequency participants may inhibit [125, 159] and enable participation at an online policy discussion [85, 165]. There are several ways to study the type of contributions that high frequency participants tend to make and how their contributions affect the participation of others. These include comparing what participants grouped by frequency contribute to a discussion [33, 15], profiling specific participants [85, 165], and experimentally varying a participant's behavior to examine whether specific behaviors are likely to elicit a response from others [212]. In some cases, it is possible to consider the type of discussion content that participants were exposed to when they enter [178, 220, 232] or the sum of what they were exposed to during their participation in a policy discussion [209].

At online community systems, like Wikipedia, high frequency participants often lead critical functions, such as welcoming, listening to, and training newcomers [29, 66, 116]. These roles are in stark contrast to the (somewhat) demonization of high frequency contributors here, as the analysis found that research about participant frequency is often to investigate issues related to incivility. One possible driver of uncivil behavior is that the useful tasks that a participant might contribute may be less apparent in the design of online discussion systems [34, 43] than in online community systems [117, 183].

Deliberation analysis approaches, such as MDC and DQI, may offer a source of inspiration for the design of useful policy discussion tasks. For example, an analysis approach developed by John Gastil and Laura Black [75, 76] was a key source of inspiration for an augmented-reality deliberation system, "provide[ing] a productive framework from which to think about design" [83, pg. 78]. With respect to MDC, it is not hard to imagine task workflows intended to promote coherent disagreement or meta-talk in a policy discussion.

Despite an emphasis on equality in deliberation literature [21, 109, 192], it is worth noting that moderators contribute to a deliberation frequently, often to prevent participants and perspectives from dominating the discussion [145, 160]. If moderator activity is included when calculating the equality of access during a deliberation, the distribution of participation will likely be skewed toward frequent contributors, i.e., less egalitarian.

We might consider alternative ways to estimate equality in online policy discussion. For example, we might separate contributions by newcomers and active participants, so as to consider what new perspectives newcomer's add to the diversity of perspective and the extent to which active participants incorporate new perspectives into a discussion. When a moderator is not available, an online policy discussion depends on its participants to engage with topics, perspectives, and each other through the discussion system.

2.5 General Discussion and Limitations

The four vignettes reflect observations based on an analysis of prevalent concepts identified through the review of online policy discussion research that references *Measuring Deliberations Content* (MDC) [207]. The vignettes raise a series of analytic decisions involved with studying reasoned opinion expression, disagreement, incivility, and the influence of high frequency contributors on the equality of access and perspective during an online policy discussion. Through the course of authoring the vignettes, I observed three types of analytic decision-making, related to the (1) dependent variables, (2) independent and control variables, and (3) theoretical concepts.

In experimental research, concepts are commonly described in dependent and independent relationships [107]. A dependent variable is expected to vary in response to some combination of independent variables. As the purpose of an experiment is to evaluate a specific set of relationships, external factors that may influence the dependent or independent variables need to be accounted for (or "controlled"). For example, a researcher will control for the *policy topic* and *time period* when comparing the reasoned opinion expression at separate online platforms [70, 94, 189]. In this example, the common discourse architectural features of each *online platform* are independent variables used to explain variation in *reasoned opinion expression* (a dependent variable).

Analytic decision-making related to the theoretical concepts reflect a more fundamental challenge involved with this research. Regardless of whether MDC was applied as part of the research methods or merely referenced in the background sections of an article, each study in this review grapples with questions about what concepts to apply and how those concepts should be operationalized. The assumptions involved with online policy discussion can be quite different than those common in deliberation, which means that applying deliberativeness as a lens with which to understand online policy discussion naturally involves some theoretical work to bridge the gaps between these different contexts for policy discussion.

The following sections review observations from the vignettes in terms of the common analytic decision-making related to dependent variables, independent and control variables, and the theoretical concepts that guide an analysis.

2.5.1 Dependent variables

The MDC deliberative elements are common dependent variables in online policy discussion research. The analysis found that few of the deliberative elements are applied to online policy discussion research, the two most prominent in the review being *reasoned opinion expression* and *disagreement*. The analytic decision-making described in this section reflect common reasons why the definition of a deliberative concept can shift in the process of determining how the concept might be operationalized for an online policy discussion context.

In the findings, I discuss how deliberative concepts are adapted at the operational-level in several ways, such as by aggregating micro-level observations into larger units. For example, in Vignette II (section 2.4.3) I discuss how aggregating thought-level observations of disagreement [158, 207] into reply-and-response [204, 211] or even network-level participant interactions [14, 156] shifts the focus of an analysis from what participants disagree with to how participants take part in disagreement. In this way, leveling a deliberative concept up and down operationally introduces different methods of interpreting and theorizing about the concept.

I also note several seemingly minor inconsistencies in the operational definitions for a deliberative concept. Recall that deliberative concepts are composed of sub-concepts: e.g., a reasoned opinion expression includes an *expressed opinion* and a *source* of evidence to justify the opinion. In some cases, slight modification to a sub-concept may have a minimal impact on observations of the deliberative concept. For example, in Vignette I (section 2.4.2) I note that research about reasoned opinion expression commonly finds that the majority of comments do not include justification, regardless of whether sourcing is operationalized as a binary or categorical variable. This may not be the case for concepts with a wide range of dimensions, such as *topic controversy*, which researchers have operationalized with various terms (e.g., sensitivity, seriousness, danger).

By comparison to an in-person deliberation, online discussion systems offer more and less visibility into the social activity of a policy discussion (see section 2.4.4). Systems are able to record how participants interact with system features (e.g., comment rating [33, 168], position polling [62, 82, 159], argument maps [86], collaborative visualizations [112, 83]), which can offer insights about participant and group behaviors, beyond what content people contribute to the policy discussion [213].

System design decisions also affect how deliberative concepts are represented and elicited in online policy discussion. When people reply to a comment posted by another participant, the reply might be perceived as more public if it is part of a hash-tag [70] or if also broadcast to a network of social media followers [189, 94], than if posted to a newspaper website. As another example, when an external source is introduced, the source might be more likely to elicit an informed discussion if participants are encouraged to raise questions about [159], annotate [6, 112], or integrate the source into their own arguments [83, 86] than if the source is merely referenced as a hyperlink in the text of a comment.

The following analytic decisions relate to how deliberative concepts are defined and operationalized as dependent variables:

- At what levels of analysis should a deliberative concept be observed?
- What terms are commonly used to operationalize a deliberative concept?

- Given access to participants, which deliberative concepts cannot be operationalized?
- What system designs are meant to represent or elicit a deliberative concept?

2.5.2 Independent and control variables

The purpose of this section is to discuss the decisions that researchers make in order to analyze how a deliberative concept relates to other factors and assumptions that might be viewed as independent or control variables in an analysis.

Deliberation is commonly used to facilitate public engagement during brief moments in the history of a policy. A shift in the state of a policy issue can affect the course of a policy discussion. For example, juries are often sequestered from information about how their case is unfolding in the court of public opinion to control for the exogenous effects that a shift in the state of an issue might have on deliberation during court proceedings [206]. The period for an online policy discussion may similarly center on specific moments of a policy history, such as when people are still exploring a shared concern [83, 82, 128]; however, often the period for discussion includes multiple shifts in the state of an issue, if not the entire cycle of decision making [14, 62, 82, 159].

Characteristics about the topics of discussion will also play into the norms of discussion. For example, the analysis found several studies that investigate how the level of controversy surrounding a policy issue, topic, or perspective affects deliberative characteristics of the discussion. In some cases, the topic is varied experimentally to examine how different levels of controversy affect discussion [8, 220, 231, 232]. More often the policy topic is controlled for in analysis, whether by selecting specific discussions to analyze [70, 156, 189] or when comparing the effects related to other discussion factors [31, 69, 94, 178, 212].

Factors related to the discussion setting and facilitation procedures also affect observations of a deliberative concept. For example, in Vignette III (section 2.4.4) I discuss research about the relationship between participant anonymity and the presence of incivility in policy discussion [15, 69, 94]. The analysis found that anonymity is commonly treated as a binary condition related to whether participants are allowed to use a pseudonym during the policy discussion or not. In the context of social media, anonymity refers to whether the comments posted to a policy discussion are review-*able* by participants in the discussion, but also by the audience of people who follow the discussion participants [70, 94, 189] (also referred to as *publicness* [198]).

The presentation of issues, topics, and existing perspectives can also influence contributions to an online policy discussion. For example, a common discussion system design decision is how to curate the thread of existing comments. In a synchronous discussion (e.g., [211, 205]), comments are typically presented chronologically, reflecting the order that the participants engaged with the policy topics [208]. By contrast, comments to an asynchronous discussion may be curated in a variety of ways in addition to chronological order [124, 154]. A few studies in the review present how such factors related to the presentation of discussion content influence participant perceptions of [209] and contributions to the discussion [178, 232].

The following are common decisions that help to identify factors to account for and to investigate as control and independent variables in an analysis:

- What events in the history of the policy does the discussion reflect?
- What topics does the discussion include?
- What social activities are transparent at the discussion system?
- What existing discussion content does the system direct participants toward?

2.5.3 Theoretical concepts

The literature review focused on articles that reference MDC [207], but I found that nearly as many articles incorporate MDC in the methods as those that do not (see Tables 2.2-2.3). For example, new concepts have been introduced to investigate incivility (section 2.4.4) and the influence that frequent contributors have on a discussion (section 2.4.5). This analysis practice of introducing concepts drawn from observations in the practice of online policy discussion is similar to how deliberation research integrates concepts drawn from the practice of deliberation (see section 2.2.2).

The purpose of this section is to discuss the analytic decision-making involved with bridging the gaps between theoretical concepts and online policy discussion. The theoretical role played by "group processes" during an online policy discussion offers a useful example. The analysis found 18 articles that conceptualize deliberation as a group activity. Examples include group decision-making [14, 227], collective action [2, 128], nominal discussion groups [31, 142, 211], and groups that gather then disband naturally during a public engagement [83, 159]. The articles that do not conceptualize deliberation as a group activity discuss research in terms of the *public sphere* proposed by Jürgen Habermas [15, 31, 94, 125, 165, 204, 205, 228] or in terms of the personal benefits that people might gain through policy discussion [211, 212, 220, 231, 232].

Articles presented in this review not only regard deliberation as a group activity, they also invoke conceptual models of human behavior in groups to explain observations in online policy discussion. In the findings, I allude to and reference several models of human behavior in groups, such as group polarization [215], group socialization [161, 162, 131], and the social identity model of deindividuation (SIDE) [186]. When investigating a social behavior that has emerged in a novel context, like online policy discussion, it is often useful to consider this behavior through the analytic lens of a discipline with more established theory and methods, such as deliberation [13] or human behavior in groups [18, 149, 130].

However, online policy discussion can involve assumptions about policy discussion that are fairly different than policy discussion in other contexts. For example, through Vignette IV (section 2.4.5) I discuss assumptions about group membership and equality in online policy discussion by comparison to deliberation. As a result, some analyses common in deliberation do not easily translate to online policy discussion, such as group polarization (section 2.4.3). When determining how to apply concepts to online policy discussion, I recommend that researchers consider how the discussion contexts differ in terms of key assumptions about group membership, facilitation procedures, and the discussion setting.

While an analysis may introduce new concepts, it is worth noting that assumptions about policy discussion are also implicit in the design of a deliberation analysis approach. Each of the deliberation analysis approaches highlighted in the related work respond to specific deliberation contexts: e.g., small group problem-solving [79, 207], national issues forums [190], parliamentary debate [203], and large multi-day deliberation events [57, 95]. Black et al. [13] recommend that researchers consider these assumptions when determining how to study a deliberation, recognizing that each analysis approach has a "compelling advantage" for some purposes [13, pg. 29]. I encourage online policy discussion researchers to do the same.

I found the following analytic decisions useful when thinking about how to navigate the conceptual gaps between online policy discussion and other contexts for policy discussion, such as deliberation:

- Are participants instructed to deliberate by themselves or collectively with others?
- If collectively, do participants identify themselves and others as members of a group?
- If as a group, what tasks are the group members expected to perform?
- What participant or collective behaviors was the discussion system intended to elicit?

2.5.4 Limitations

The analysis of analytic decisions for studying deliberativeness is not without limitation. First, I chose to center our review on the articles that reference *Mea-*

suring Deliberation's Content (MDC) [207], where I could have selected another approach, like the Discourse Quality Index (DQI) [203]. MDC has several properties that I feel are useful for thinking about online policy discussion (see section 2.3.1), but by narrowly focusing on MDC I may have missed some decisions that are common when applying another approach.

Second, I decided to conduct a literature review, rather than interview experts or organize a focus group (or deliberation) to discuss the analytic decisions that researchers commonly make in their practice of studying online policy discussion. I also chose an approach that involves a communication-based hand-coding analysis of deliberativeness, rather than a computational or natural language processing (NLP) based approach.

Third, the findings are based on an in-depth analysis of a subset of the concepts we identified through the review process. While every article played a part in the vignettes, the focus on prominent concepts means that I may have missed some analytic decision-making related to concepts that were less prominent.

It is also important to note that I apply MDC [207] not as a gold-standard for online policy discussion or deliberation analysis, nor to characterize the research that deviates from MDC as erroneous. Rather I use MDC as a means to investigate common analytic decisions in the practice of studying deliberativeness in online policy discussion.

2.6 Conclusion

In the practice of research, it is common to draw inspiration from multiple disciplines. In this chapter, I examine how concepts from the theory and practice of deliberation have informed the analysis of online policy discussion. However, the process of applying concepts from deliberation is not straightforward. First, there are many theories and practices related to deliberation [170, 218], which can be challenging for researchers to conceptually navigate (Medaglia and Yang [156] offer an illustrative example). Second, deliberation typically involves a small group of people who meet in person for a brief period of time to consider a policy scenario through a moderated discussion of diverse perspectives [21]. These conditions are different in online discussion [32, 72].

To shed light on the practice of studying deliberativeness in online policy discussion, the chapter presented a systematic review of the research referencing one specific analysis approach, i.e., Jennifer Stromer-Galley's *Measuring Deliberation's Content* [207] (MDC). The chapter reviews how concepts from MDC, like *reasoned opinion expression*, have been applied (section 2.4.2) and adapted (section 2.4.3), and how researchers incorporate new concepts (section 2.4.4) and assumptions to study deliberativeness in online policy discussion (section 2.4.5). The chapter presents a synthesis of the observations from each vignette into a set of common analytic decisions. The common analytic decisions might serve as an initial guide when preparing to study deliberativeness or when studying existing research about deliberativeness in online policy discussion.

The common analytic decisions developed through the course of this chapter serve as a conceptual foundation for the remainder of the dissertation. In Chapter 3, I introduce the Cornell e-Rulemaking Initiative (CeRI) RegulationRoom platform, by reviewing how design choices related to the discourse architecture engage with assumptions about policy discussion in terms of the group membership, facilitation procedures, and setting for discussion. The case studies, presented in Chapters 5 and 6, demonstrate how to apply deliberative concepts to analyze and design for system features in an online policy discussion. Together these components of the dissertation contribute to a central argument that the challenges involved with studying deliberativeness in online policy discussion, highlighted by this chapter, can be addressed by tightly integrating analysis with system design research.

CHAPTER 3

DISCOURSE ARCHITECTURAL DESIGN CHOICES

Policy touches nearly every aspect of our daily lives, from local traffic laws to the privacy provisions of Facebook and other online platforms. It is challenging for decision-makers to understand the broad human experience of a policy, which is why public engagement is particularly important for policy making. To quote John Dewey's [47] argument for direct public engagement with public matters, "the man who wears the shoe knows best that it pinches and where it pinches, even if the expert shoemaker is the best judge of how the trouble is to be remedied" [47, pg. 364].

Discussion is central to public engagement. For example, public meetings typically involve a public commenting period for people to raise perspectives for policymakers to consider (e.g., Robert's Rules of Order [187]). As another example, design charrette is a common practice in local/regional planning, where members of the public are invited to learn about and discuss their impressions of a proposed development (e.g., bikeways, changes to a transit network) [217]. While public commenting and design charrettes offer people an opportunity to add their perspective to a policy discussion, design charrettes are organized to elicit informed perspectives in a way that public commenting processes typically are not.

Deliberation offers yet another model of public engagement. As introduced through Chapter 2, the stages of a deliberation introduce people to specific issues related to a policy, which is similar to the educational components of a design charrette. However, unlike a design charrette, deliberation is used to identify and explore opportunities for disagreement, by challenging participants to carefully weigh their perspectives against perspectives that differ from their own [78, 108]. Participation in a deliberation is also tightly controlled—participants are recruited and assigned to deliberate, where public commenting is by definition open to the public [143, 217].

As the objectives for these types of public engagement differ, organizers of these events use the discussion setting and facilitation procedures to reinforce power relationships among group members [137, 174, 175]. For example, Olson [174] describe how the seating arrangement for a house of parliament typically positions opposing political parties directly opposite each other to provoke debate. As another example, during a public commenting period the public audience is typically seated below the policy makers, with just one or two microphones to share and only a few minutes per person to speak. Design charrettes and deliberation offer a stark contrast, as organizers will prepare their physical space to promote equality among participants. At a deliberation, participants might meet around a common table to discuss evidence about a shared concern [78, 108, 217].

As described in Chapter 2, the concept of "Discourse Architecture" is used to discuss how the design of a public space influences the norms of discussion that are likely to emerge within the space [70, 225]. Decisions about a discourse architecture play into assumptions about group membership, facilitation procedures, and discussion setting. The following are several questions intended to prompt conversation about how design choices related to a discourse architecture affect assumptions about policy discussion:

- *Group membership*: Who can participate, how do people relate to each other, and what roles do people perform for the group?

- *Facilitation procedures*: How are participants exposed to the topics, perspectives, and other group members within the discussion setting?
- *Discussion setting*: How does the public space allocate support, power, and respect among the group members?

As with public engagement in a physical setting, system design does not play a neutral role in online discussion about policy [225]. While any online discussion system can be used to host an online policy discussion—from synchronous chat to the commenting interface at a newspaper website—the case studies presented in Chapters 5 and 6 involve a discourse architecture modeled after the Cornell e-Rulemaking Initiative (CeRI) RegulationRoom. The RegulationRoom platform offers an interesting context for thinking about the interplay between design and analysis, as the discourse architecture incorporates some features that are common in discussion forums and some features that make it similar to a deliberation system.

The section that immediately follows serves to introduce the discourse architecture of the RegulationRoom in the context of online discussion and online deliberation systems. The RegulationRoom system and facilitation procedures are discussed in terms of the above assumptions about group membership, facilitation procedures, and the setting for discussion. Chapter 4 introduces the specific case studies presented by the dissertation, emphasizing how assumptions about policy discussion in the RegulationRoom discourse architecture were adjusted to analyze and promote specific deliberative concepts through the design research and development presented in Chapters 5 and 6.

3.1 A blurry line between discussion and deliberation systems

The RegulationRoom platform occupies an interesting design space between online discussion and online deliberation systems. Online discussion and deliberation systems are similar in several respects. The recommended best practices for the design of online deliberation systems [34, 40, 43, 219] include many recommendations that also appear in the best practices for designing online discussion community systems [117]. As these online systems aim to facilitate conversation among groups of people, similar attributes of the communication medium also affect how people relate to the policy topics, perspectives, and each other at each type of system [32, 72].

While deliberation and online discussion systems share many properties, the remainder of this section focuses on their differences, so as to set up a later discussion about how the RegulationRoom discourse architecture sits on a blurry line between these types of system. There are two key factors that distinguish deliberation systems from online discussion systems: Deliberation systems support *collective decision-making* to advance *specific outcomes*.

In order to support collective decision-making, a deliberation system presents information about a policy issue, provides mechanisms to facilitate analysis of the information, and enables participants to prioritize insights developed through analysis. Deliberation systems offer access to informational resources (e.g., reports, news, plans, video), encouraging an informed discussion about specific policy issues [40, 219]. For example, before joining a deliberation during the Virtual Agora Project all participants were introduced to the issues for deliberation through a "library session" [163, 164].

Deliberation systems offer access to information, but also mechanisms to promote analysis of that information. Deliberation systems typically include features to support discussion [11, 61, 185]. Many deliberation systems also integrate a contribution protocol to elicit participant information in forms that are amenable to computational representation, such as the Issue-Based Information System (IBIS) [122]. The information structure elicited by a contribution protocol can be easily searched [36, 200], visualized [46, 115], or represented as pro-con arguments [118] or as a layer of information over a budget document to highlight the social issues underlying key line items [112].

However, there are limitations related to a contribution protocol, like IBIS. While standard discussion systems are fairly familiar for people, contribution protocols can be tricky to learn and therefore present a barrier for less technologically savvy participants. On the other hand, a representation of existing contributions, as a pro-con list [118] or budget data visualization [112], may offer a lower barrier for participants to become informed about existing topics and perspectives than by reading a bunch of comments. Additionally, the logged activity of user interaction with a representation of existing contributions, such as an argument map [46, 115], may reveal insights about how people engage with the policy materials, perspectives, and other stakeholders in the issue.

After exploring and interrogating the informational resources at a deliberation system, participants are encouraged to prioritize their insights through a decision-making process, so as to advance specific outcomes related to the policy issues [34, 40, 43, 219]. However, "decision-making process" can imply many things. Local governments often adopt Robert's Rules of Order as a process to facilitate decision-making during public meetings [187]. Robert's

Rules center around a meeting agenda of issues for discussion and voting procedures to negotiate the issues as well as the agenda. Several deliberation systems have adopted Robert's Rules to facilitate asynchronous discussion and decision-making [44, 195, 219].

However, the process of a deliberation might not be staged or negotiated. As an example, the *ConsiderIt* platform continuously generates pro-con lists that reflect alternate perspectives of a policy issue, by coordinating information about how people add and borrow arguments to craft pro-con lists that reflect their own perspective [118]. As people use ConsiderIt, the system gains information about the mix of arguments that shape participant perspectives as well as the level of popularity and controversy associated with each argument. Where Robert's Rules offer a process for humans to deliberate with each other, ConsiderIt coordinates human input to computationally map arguments to a range of perspectives about a policy issue, which resembles deliberation.

While an online discussion may continue indefinitely, deliberation is intended to produce specific outcomes related to a policy issue [34, 40, 43, 219]. For deliberation to be relevant—whether to inform a policy decision, to promote public education, or to form community around a shared concern—the time line associated with a collective decision-making process must also align with the opportunity for the deliberated outcomes to influence the policy issues [34]. When the opportunity for influence is narrow, an organizer might adopt systems that use voting procedures and rational argument to rapidly generate specific outcomes [1, 115, 200]. However, to affect a long range societal concern an organizer might adopt a system with features that encourage participants to form a collective identity with the issues through an online deliberation community [34].

These fundamental questions about the process for *collective decision-making* and the *specific outcomes* intended for deliberation are what make deliberation systems different from online discussion. Design choices in the RegulationRoom discourse architecture position it along the blurry line between online discussion and deliberation systems design. As a platform originally intended to seed discussion from informed public comments, RegulationRoom resembles a standard discussion forum, yet through a multi-staged process of discussion and synthesis the RegulationRoom is a resource for the CeRI team to advance specific outcomes, i.e., insights to assist US federal government rulemaking.

3.2 Case study: The RegulationRoom discourse architecture

The US federal rulemaking process consists of three stages: pre rule, proposed rule, final rule. During the proposed rule period, federal agencies are required to solicit public feedback for at least 30-60 days. After the comment period closes the federal agency will review all of the comments and either remove, modify, or finalize the proposed rule. Final rules become law no less than 30 days after the comment period ends—a brief opportunity for public influence. In October 2002, the US eRulemaking Program was created as a cross-agency initiative under Section 206 of the 2002 E-Government Act (H.R. 2458/S. 803). In 2003, the eRulemaking Program launched an online system, called Regulations.gov, to increase the transparency of the federal rulemaking process and facilitate online public commenting.

The CeRI RegulationRoom was an experimental online public learning participation platform that was also designed to facilitate online public commenting about US federal rulemaking, but RegulationRoom was intentionally designed to elicit informed perspectives, referred to as *situated knowledge*. R. Farina et al. [185] define situated knowledge as, "information about impacts, problems, enforceability, contributory causes, unintended consequences, etc. that is known by the commenter because of the lived experience in the complex reality into which the proposed regulation would be introduced" [185, pg. 148]. Situated knowledge is often conveyed through stories rather than direct argumentation [60]. Policymakers often need this type of insight as they weigh the different options to a social policy problem [47].

As is common when preparing for either an online policy discussion or an online deliberation, the CeRI team carefully constructed background material about the proposed changes to US federal rules before each public commenting period. It can be challenging to interpret the legalese associated with a policy. People have different levels of skill, training, and time available to study a policy and also to consider their own feelings about the matter [192]. To reduce this common barrier to the public's engagement with policy issues [47], the CeRI team worked with the federal agencies involved to translate and then triage a proposed rule into a set of 6-8 policy topics where public input would be meaningful to the federal agency proposing the regulatory change.

In the process of translating and triaging the rule so that it would be accessible to a layperson audience, the CeRI team also used hyperlinks to layer each summary with multiple levels of information about the policy proposal. Some hyperlinks integrated a hover-over message effect to provide participants with easy access to simple information, such as a definition or a brief synopsis of a policy process. Hyperlinks were also used to direct participants through a sequence of related topic summaries and even toward specific sections of the proposed rule itself. By translating and triaging the policy into topic summaries and then layering each summary with information, participants at RegulationRoom could choose their own level of detail into the background materials.

The background materials were used to structure the online discussions. To promote an informed policy discussion about each of the 6-8 policy topics, the topic summaries were positioned in RegulationRoom immediately to the left of a public commenting interface specific for each topic and subtopic, in a two column and multi-page layout [185]. This design segments the online discussion into designated online spaces for each topic, but also requires that participants determine the "right" space to post their comment.

Similar to many online discussion systems, RegulationRoom incorporated some methods for filtering the discussion. When a discussion includes a large volume of existing comments it can be challenging for people to make sense of the topics and to identify an opportunity to contribute to the discussion [42, 115, 123]. Two content filtering strategies were implemented at RegulationRoom: The first filter reduced the discussion to comments posted by a specific participant and the second reduced the discussion to comments "recommended" by the moderator team. However, it is less clear if or how participants may have used these filters.

The CeRI moderator team was fundamental to the experience of RegulationRoom. CeRI recruited law school students through an innovative e-Government Clinic that exposed students to the theory and practice of conflict resolution and collaborative decision-making online, while also training the students in "active listening" techniques and offering a hands-on experience moderating the RegulationRoom discussions. As the purpose for engaging with commenters was to encourage "more, better participation through informed and substantiated arguments," moderators invested time to understand each comment in the context of what a commenter is trying to convey to other commenters [61, pg. 14].

Every comment added to the site received moderator attention, which meant "reading the comment and the surrounding context, deciding whether and how to respond, drafting a response, and emailing back and forth with supervisors until their response was approved and could be posted to the site" [61, pg. 10]. Thousands of people visited the RegulationRoom to contribute to the online policy discussion of several proposed rules; however, the vast majority of participants participated just briefly. Even in the most popular consultation, contributor median visit duration was one day, and over one-third of all comments were made by one-timers.

A comment posted by a one-time participant might offer an informed perspective that advances the discussion, but the comment also imposes work for the moderator team to draft, finalize, and get approval to send a reply that will never actually see a response from the commenter. Comments posted by newcomers to an online policy discussion are also more likely to be off-topic than comments posted by regulars [33, 94]. As a result, even a well crafted one-time comment that contributes an informed perspective to the discussion, might also distract attention away from the other topics under discussion.

The RegulationRoom process for online public commenting incorporated an "evolving mix of human, automated, and computer assisted elements" to foster an informed online policy discussion, but also to yield outcomes similar to those generated by deliberation systems [185, pg. 396]. In large part, the RegulationRoom experience depended on the moderator team to manage the stream of comments and engage with commenters as well as to synthesize the discussion content into policy recommendations suitable for use by the US federal agency partners. This substantial human effort is what made the RegulationRoom experience hard to replicate in other policy contexts and hard to scale out to larger audiences [61].

When I joined the CeRI team, in the summer of 2013, my first project was to investigate how the discourse architecture of RegulationRoom might be modified to elicit better contributions from newcomers to the discussion. An improvement to the quality of newcomer contributions would reduce the number of participants who the moderator team would need to help articulate their situated knowledge of the policy. The following sections introduce the research procedures and policy context that I developed to investigate how factors related to the discourse architecture of RegulationRoom affect the quality of first-time contributions, which I defined as a comment's "Responsiveness" to the policy topics.

3.3 Ethical considerations for discourse architecture experiments

Disenfranchisement means depriving a person of their legal rights. In a democracy these legal rights include the right to vote and to voice opinions about a policy issue. It is also valuable for a democracy to encourage people to become informed about policy issues, as a way to promote feelings of citizenship, but also to solicit public input that informs policy making [21, 25, 47, 55, 87]. For these reasons, there is democratic value in discourse architectures, such as that of RegulationRoom, which are intended to promote informed online discussion about policy [185].

It is often hard to detect when a discourse architecture has disadvantaged a group of the stakeholders involved in a policy discussion. If a particular discourse architecture exaggerated the dominance of one perspective over another, for example, then the people opposed to the exaggerated perspective might feel less willing to voice their opinion or even to join in the discussion—referred to as the "spiral of silence" theory of public opinion [172]. However, without asking people to reflect on their commenting practices or conducting controlled discourse architectural experiments, it is hard to infer what factors affect the content that participants contribute to a policy discussion.

As RegulationRoom was used to facilitate public commenting about proposed changes to US federal rules, the CeRI team did not conduct controlled experiments during live rulemaking periods because of the concern that a condition might adversely affect, if not disenfranchise, people hoping to voice their opinion about an issue. To investigate factors of the RegulationRoom discourse architecture that affect newcomer contributions, I developed a research procedure that involved recruiting Amazon Mechanical Turk (AMT) crowd workers (called "Turkers") as part of controlled experiments to participate in an online discussion about the AMT Participation Agreement at a RegulationRoom-like platform [153, 154, 155]. I chose this population and topic because it created a policy context where participants would have interest in, experience with, and opinions about the subject matter, much like the people who are likely to participate in online public commenting.

AMT is an online crowd labor market, where people can find, accept, and complete tasks for a monetary reward. The AMT platform facilitates the labor relationship between the people (and systems) who request tasks (called “Requesters”) and the Turkers who perform tasks. In this way, a layer of technology separates Amazon’s Turkers from the Requesters for whom they complete work. This separation makes it possible for Requesters to coordinate large crowd workforces, but it also means that each transaction with a worker is mostly anonymous, abstract, and legally ambiguous [152]. These conditions raise concerns about fairness and abuse [103, 147].

These concerns are exacerbated by AMT’s hands-off approach to the labor market, which was in effect from 2005-2018. AMT’s participation agreement classified Turkers as independent contractors, free to accept any task they qualify for. At the same time, Requesters have the right to reject a Turker’s completed work without payment while AMT, providing only the venue for an exchange, is not involved in resolving any labor disputes. When a Turker’s work is rejected, the result is lost pay, time, and reputation, and AMT’s stance gives workers little recourse. While many rejections may be warranted, prior to November 2018 Turkers had no formal way to report rejected work that they felt was rejected unfairly, which made the experience of these rejections particularly troubling.

While Turkers by-in-large acknowledge the risks of rejected work at AMT, they disagree about the necessary steps to amend the AMT participation agreement [152]. A particularly controversial proposal is that when work is rejected by a Requester, the Turker would be paid for the parts of the work that are acceptable or still useable by the Requester (called “partial payment”) [154, 155]. On the one hand, offering partial payment acknowledges the time and effort that Turkers commit to perform a task. On the other hand, offering partial payment may have broad implications for the crowd labor market. Partial payment might entice Requesters to automatically reject Turker work in order to pay less or might motivate Turker's to perform low quality work, as any quality of work will yield some minimum.

The experiments presented in Chapters 4 and 5 use the controversy of the partial payment proposal to examine how participants respond when exposed to or asked to perform a task that involves perspectives that are different than their own. However, unlike traditional labor markets, it is not clear how to address crowd work labor disputes through existing regulatory authorities, and unlike other online platforms, such as Reddit [24], Turkers are not well positioned to effect change in AMT. Performing research in this experimental context therefore has special ethical circumstances that need to be considered.

At the individual level, academic researchers can (and have) adopted practices that are intended to reduce Turker-experienced risks associated with crowd work. These practices include receiving IRB-approved informed consent from all participants and compensating participants for their time based on Turker approved standards for academic research. There are also several Turker-supported best practices for crowd work task design and some Requesters have recruited Turkers to help improve and error test a crowd work task before releasing it [151, 191].

To address these concerns through the research presented in this dissertation, I took the following steps when working with Turkers on tasks that involved the AMT policy context:

1. All human intelligence task (HIT) related work was conducted on the AMT platform as opposed to through a HIT that requires Turkers to access an external website. While the AMT participation agreement requires that HITs are performed on the AMT platform, in practice this guideline is regularly ignored.
2. The reward for all HITs related to the work presented in this dissertation was priced to reflect the New York state minimum wage at the time of HITs release.
3. Whenever a batch of HITs related to an experiment were released to the AMT marketplace, another dummy HIT was simultaneously posted to compensate participants for any technical errors or concerns related to their participation in an experiment.
4. I personally responded to every participant inquiry and actively monitored Turkopticon and TurkerNation to respond to any participant concerns posted to those forums.
5. I received informed consent from every participant in each experiment.

3.4 Choosing to study a deliberative concept

It took about a year of design, development, and testing to conduct my first (successful) experiment with AMT Turkers in a RegulationRoom-*like* discussion about issues related to the AMT Participation Agreement [151]. The initial experiment investigated how factors related to the phrasing of message prompts in the discussion affect the *Responsiveness* of comments posted by first-time participants. I found that even subtle details, like the content of the default message in a comment text box, can yield contributions that respond to (or distract attention away from) the policy topic of the discussion [153].

Note in the prior paragraph that "Responsiveness" is mentioned as an important dependent variable meant to capture whether contributions respond to or distract attention away from the policy topic of the discussion. Much like the absence of "incivility" in Jennifer Stromer-Galley's [207] *Measuring Deliberations Content* (MDC), a simple word search will confirm that the term "responsiveness" is not mentioned in the text of Chapter 2, nor is it mentioned in any other approach used to analyze the content of a deliberation (to my knowledge). Instead the term, "topic coherence" is more commonly used to evaluate what I had conceptualized as responsiveness.

Our team developed the concept of responsiveness to reflect the hierarchy associated with the presentation of the policy background material in RegulationRoom. As mentioned previously, prior to each commenting period the CeRI team worked with US federal agency partners to translate and triage the policy material into issues, topics, subtopics, and discussion prompts. This hierarchical structure was designed into the presentation of background material and the position of the discussion interface within RegulationRoom. Specifically, the concept that we used for analysis was inspired by the discourse architecture that we analysed.

As the RegulationRoom process for online public commenting is somewhere near the boundary, but not entirely a deliberation system the deliberative concept of topic coherence was foreign at the time. However, even with the clarity of hindsight, our operationalization of responsiveness offers a more granular level of insight into the questions that guided our research than the standard definition of topic coherence. Yet, for a long time after publishing our research I did wonder about our decision to create a new concept, when we might otherwise have added scholarly work to an established alternative, i.e., topic coherence.

The conceptual dilemma associated with this selection of our dependent measure initially peaked my interest into the interplay between analysis and design when studying the deliberativeness of an online policy discussion. Where “responsiveness” was conceptualized for the RegulationRoom discourse architecture, adopting “topic coherence” as a dependent measure would require some adaptation in order to accommodate the different assumptions about discussion at RegulationRoom and discussion during a policy deliberation. The case study presented in Chapter 5 investigates these considerations by way of an experiment about the effect of content curation decisions on the topic coherence of newcomer contributions to an online policy discussion.

My experience crafting the concept of responsiveness also helped me to recognize the role that design research might play in the process of adapting deliberation concepts to explain the social behaviors that emerge in online policy discussion. Chapter 6 develops and evaluates a process for onboarding newcomers to an online policy discussion, so as to investigate tactics to elicit metatalk about points of conflict. Together the case studies contribute insight about the interplay between design and analysis when determining how to apply concepts from one social context to study another.

While my connection with the CeRI team is what set me down this path, the discourse architecture of RegulationRoom offers an interesting case to examine what it means to adjust the deliberativeness of an online policy discussion through analysis and design. As the CeRI team heavily supported the deliberation aspects of the RegulationRoom process, the discourse architecture of RegulationRoom without the CeRI team is more similar to a discussion system than it is a deliberation system. To investigate the process of analyzing and designing for deliberative concepts, I decided to adopt the following assumptions about online policy discussion from RegulationRoom's discourse architecture through the experiments presented in Chapters 5 and 6:

- *Discussion setting*: In each experiment, the policy context was triaged and translated to be accessible for a layperson audience, but also layered with links to direct interested participants toward specific policy sections. Like RegulationRoom, the policy materials for the experiments were presented adjacent to the online discussion interface in a two-column design, so as to promote an informed discussion.
- *Facilitation procedures*: The most critical differences between the experiment protocol and RegulationRoom relate to the facilitation procedures. Where a team of moderators actively facilitated the online discussion at RegulationRoom, the experiments in Chapters 5 and 6 examine how the system design might elicit specific types of contribution to the discussion.
- *Group membership*: While targeted recruitment was used to identify participants from a specific group of stakeholders, for experimental purposes the online discussions during the experiments were not available to the general public. Similar to RegulationRoom the participants communicated with pseudonyms, though the naming convention for usernames

was standardized, i.e., concatenating a random color name with a type of animal.

CHAPTER 4

INTRODUCTION TO THE CASE STUDIES

The case studies in Chapters 5 and 6 illustrate the interplay between analysis and design with deliberative concepts. Both studies were published at academic conferences and are presented in the dissertation unaltered from their published form, so as to demonstrate how my own writing about and use of deliberative concepts in research and design has shifted over time.

Two concepts from Jennifer Stromer-Galley's [207] Measuring Deliberations Content (MDC) guide the research presented in the case studies: i.e., topic coherence and meta-talk. In a discussion, topics advance as people reply and respond to each other along a common thread of subjects. Topic coherence in online policy discussion refers to the consistency of the topics within a thread of comments. When discussion participants are often off-topic, or move away from topics too quickly, the discussion cannot deeply consider a policy issue.

Meta-talk is talk about the state of a discussion. As presented in Chapter 2, common observations of meta-talk are related to the level of incivility in an online policy discussion [14, 33]. However, MDC defines three forms of meta-talk that go beyond incivility and rather relate to the problem-solving that deliberation groups perform to weigh a policy issue by addressing opportunities for conflict, consensus, or for clarification. While the ability of a group to continue discussion about a topic—high topic coherence—is necessary to deeply consider an issue, meta-talk is used to prompt further analysis into the issue.

Recall that when applying deliberative concepts, such as topic coherence or meta-talk, to study policy discussion within a novel context, researchers make some conceptual and operational adjustments to account for differences in the assumptions about policy discussion. In Chapters 2 and 3, I discuss the following assumptions about policy discussion:

1. *Group membership*: Who can participate, how do people relate to each other, and what roles do people perform for the group?
2. *Discussion setting*: How does the public space allocate support, power, and respect among the group members?
3. *Facilitation procedures*: How are participants exposed to the topics, perspectives, and other group members within the discussion setting?

In both case studies, topic coherence is a dependent measure used to evaluate the effects of system design decisions related to the facilitation procedures. In the first case study (Chapter 5), system design decisions about content curation are varied to examine the effects of prioritizing alternate positions of a controversial policy topic on the coherence of newcomer contributions to the existing topics. In the second case study (Chapter 6), a preliminary stage is added to the facilitation procedures, so that before entering an online policy discussion newcomers are introduced to the discussion content by crafting a discussion prompt based on two existing comments—performing a task inspired by meta-talk about opportunities for conflict.

The case studies involve different assumptions about facilitation procedures than assumed in the MDC. As presented in Chapter 2, MDC was initially developed to study the Virtual Agora Project deliberations. The facilitation procedures during the Virtual Agora Project involved two 90 minute sessions of synchronous policy discussion among ~8 participants over the course of a day, with breaks and opportunities to study background materials related to the policy in between sessions. During each session, participants were permitted to speak for up to 3 minutes per turn. Participants could add themselves to a speaking queue and "jump the line" up to 3 times during each session, using a synchronous audio system called PICOLA to administer their speaking order. By contrast, the experiments presented in Chapters 5 and 6 simulate the experience of a newcomer joining an asynchronous online discussion that is already underway.

The case studies also involve assumptions about facilitation procedures that are different than those assumed in RegulationRoom. While the RegulationRoom discourse architecture is applied through the case study experiments, in practice the RegulationRoom required regular attention from a team of trained moderators to monitor the stream of comments and to engage personally with each newcomer. As lamented in Chapter 3, the costs associated with moderating the platform contributed to the end of the CeRI project. The case study experiments examine system designs intended to promote deliberative properties, whether a human moderator is available or not.

Rather than treat facilitation as an assumption to consider when observing an online policy discussion, the case studies consider how the practices of facilitation affect participant contributions to the discussion—as an independent variable. The case studies investigate design levers to affect the likelihood that participant contributions to a policy discussion are deliberative, by choosing what content participants are exposed to first (Chapter 5) and by providing instruction and support for participants to craft contributions that exhibit deliberative properties (Chapter 6). Such explanatory research is important ahead of a deliberation, as there may only be one opportunity to facilitate discussion about a particular policy issue, with specific stakeholders, and perspective.

The purpose of this introduction is to explain how the case studies help to illustrate the conceptual implications to shifting the assumptions associated with a deliberative concept.

4.1 New group members in asynchronous discussion

In a synchronous discussion, all participants are able to send and receive messages at once and simultaneously [32]. In face-to-face deliberation, for example, a person listening to another can use eye contact or a smile to indicate their attention and approval, information that the person speaking receives in real-time. Additionally, the sequence of turns during a synchronous discussion tends to form a natural order as participants initiate, continue, and conclude a sequence of exchanges about particular topics [81]. The Virtual Agora Project offers a slight variation as the order that topics are discussed is, in part, due to the queue to speak.

When a synchronous discussion is recorded (e.g., as text, audio, video), the discussion may be "reviewable" meaning that participants can pause their exchange and rewind through the issues, topics, and perspectives previously raised [32]. Reviewability is less feasible in synchronous discussion, which means that the participants rely on their personal memory or notes to advance discussion of a topic. In the AmericaSpeaks deliberation protocol, for example, a group of notetaker's called the "theme-ing team" was available to assist moderators with real-time access to a synthesis of discussion topics [108].

The experiment protocol applied in the case studies, simulate a newcomer joining an asynchronous discussion that is already underway. By comparison to synchronous policy discussion, comments to an asynchronous discussion tend to be longer and are more likely to include deliberative elements, as the temporal delay associated with an asynchronous discussion can allow for a more considered response [205]. The case studies experimentally vary aspects of an online policy discussions reviewability, by changing the content curation strategy (Chapter 5) and by introducing a task that involves synthesizing content from the discussion (Chapter 6).

Shifting from synchronous to asynchronous discussion also affects the group membership. Participants in the Virtual Agora Project began simultaneously and took sequential turns contributing content to the discussion during a 90 minute period. By contrast, the participants in the case study experiments joined as newcomers to a discussion. Newcomers tend to "investigate" an online discussion by reading through the existing comments to evaluate their potential benefit to participation. As argued in Chapter 6, this preliminary investigation can yield benefits to the discussion group [29, 116]; however, the benefits that newcomers might bring to a discussion were not available for the synchronous groups organized for the Virtual Agora Project, because their membership was fixed to a set size, composition, and duration.

Where membership is typically restricted in the practice of deliberation, in an asynchronous online policy discussion the level of participation ebbs and flows. The experience of a newcomer joining a discussion is one part of that cycle, where other parts include the experience of moderators or of regular members confronted by a rising tide of newcomers. However, the design choices examined through the case studies are viable ways to influence deliberative properties of a policy discussion, because the discussion was reviewable and the participants—being newcomers—were motivated to review comments already in the discussion.

4.2 Policy discussion facilitation without a moderator

Moderators play an influential role in group deliberation [145, 190] and in online policy discussion [61]. During a deliberation, moderators help to maintain a productive social environment for the policy discussion and to keep their group focused on the objectives for the policy deliberation [145, 160].

Some deliberation analysis approaches, such as MDC, include communication coding procedures to evaluate what influence a moderator may have had on the course of a policy discussion. When a deliberation includes a strong moderator, the discussion might not reflect a group of people talking with each other, but rather separate dialogues between each member and the moderator [190]. Moderators also affect the topic coherence of the discussion. Stromer-Galley [207] specifically calls attention to the tendency of moderators during the Virtual Agora Project to ask follow up questions about comments that were not on-topic.

While moderators can play an influential role in facilitating a policy discussion, their work can be taxing [61]. For example, the RegulationRoom was limited by the capacity of its moderator team to process the stream of comments during a public commenting period. As another example, in a field trial of the *Deliberatorium* it took a team of moderators to organize and "clean up" participant contributions to the platform's collaborative argument map [86]. Gürkan et al. [86] estimate that an appropriate moderator-to-participant ratio for managing Deliberatorium would be somewhere between one moderator for every 10-20 participants, which is not far off from the average group size for in-person deliberations (see Chapter 2).

In their *Future Directions for Public Deliberation*, Levine et al. [133] call for new strategies to scale deliberation practices out to larger audiences. The case studies evaluate system design strategies that might offload some of the work performed by moderators onto the discourse architecture, by modifying how the facilitation procedures engage newcomers with existing discussion content. On the one hand, deliberation can more easily scale out to larger audiences if less of the work involved with facilitating the policy discussion is performed by humans. On the other hand, doing so raises questions about how to assess the influence of a discussion system on the allocation of topics, perspectives, and power during an online policy discussion.

4.3 Shifting from analysis to design with deliberative concepts

A common use of MDC is to compare a series of deliberations by the amount of discussion exhibiting deliberative elements [207]. For example, 66% of the thoughts expressed during the Virtual Agora Project deliberations were on topics other than the main topics structuring the discussion about a school reform initiative. While the "other" topics participants raised were not unrelated to schools, they did not address the deliberation topics. Instances of meta-talk during the deliberations were similarly summarized as a percentage of all expressed thoughts: i.e., meta-talk about conflict (0.1%), meta-talk that clarifies (0.1%), meta-talk about opportunities for consensus (0.7%).

At the discussion-level, the Virtual Agora Project deliberations were mostly off-topic and meta-talk about conflicts rarely occurred. These findings might serve a variety of purposes. For a school board member weighing the school reform initiative, the discussion-level results might discredit any policy recommendations that derive from the Virtual Agora Project, due to the low level of coherence with the main topics. For a system designer, the discussion-level results might offer a baseline to compare future advancements to the PICOLA synchronous audio system. The moderator team might also use the discussion-level results to reflect on their facilitation strategy, as Stromer-Galley [207] noted that moderators may have contributed to the low level of topic coherence.

Rather than analyze an online policy discussion after the fact, the case studies report findings from tightly controlled experiments to investigate facilitation strategies that might affect the content of participant contributions to an online policy discussion. Fundamentally this reflects a different purpose for research than the analysis of the Virtual Agora Project reported in MDC. The case studies reflect the type of analysis that the PICOLA system designer or the Virtual Agora Project moderator team might conduct to prototype new discussion facilitation strategies ahead of a deliberation. Conducting this type of user research is costly, but there may only be one opportunity to facilitate an online discussion about a particular policy issue.

However, much like the practice of deliberation, there is a wide range of possible ways to facilitate an online policy discussion. As detailed in Chapter 2, there are countless design choices: Should participants be anonymous or personally identifiable? What type of discussion should the system facilitate (e.g., debate, deliberation, dialogue, storytelling)? How might the platform represent issues, topics, and perspectives in the discussion? Should the discussion be threaded, include functions for direct reply, hashtags, private groups, user mentions? And so on. The wide range of options can make it hard to know where to start prototyping new facilitation procedures.

The case studies provide examples of how to integrate deliberative concepts into the design of new facilitation procedures. Designing a discourse architecture after a deliberative concept is similar to operationalizing a deliberative concept for analysis. Recall from Chapter 2 that an operational definition is a complete and explicit explanation of how each component of a conceptual definition will be measured. For example, the *Reflect* discussion implements the concept of "active listening" as a workflow for people to follow when replying to a comment. The Reflect workflow serves as a complete and explicit explanation of how to craft comment replies that include properties of active listening [119]. Chapter 6 develops and evaluates a similar workflow modeled after meta-talk about conflict, where newcomers are instructed to construct policy discussion prompts by synthesizing the content from existing comments to a discussion.

On their own, the case studies offer insights about the design of online policy discussion systems and provide evidence that newcomers might offer more to a discussion than comments alone. Together the case studies demonstrate the tight relationship between analysis concepts and system design. In Chapter 7, I discuss how the design "work" involved with deconstructing a discourse architecture into its conceptual components may generate new insights about where and how to bridge practices developed through policy discussion system design back to theory and practices of analysis and design in deliberation.

CHAPTER 5
CASE STUDY 1: EFFECTS OF COMMENT CURATION ON COHERENCE

5.1 Forward

With the exception of several formating adjustments, the chapter presents research published at the 2018 ACM International Conference on Supporting Group Work (GROUP) under the following citation:

McInnis, B., Cosley, D., Baumer, E. and Leshed, G., 2018, January. Effects of Comment Curation and Opposition on Coherence in Online Policy Discussion. In *Proceedings of the 2018 ACM Conference on Supporting Groupwork* (pp. 347-358). ACM.[1]

The chapter centers on the deliberative concept of *topic coherence* with the topics that emerge through participant interaction during a discussion, which is different from topic coherence with the topics that structure a discussion. The chapter highlights several analytic decisions that emerged through the process of working with topic coherence as a dependent variable.

As operationalized for the research presented in this chapter, topic coherence reflects the coherence between the topics in a newcomers' first contributions and the existing topics in an asynchronous discussion. The specific operational definition was adapted from Stromer-Galley and Martinson [208], which present research about topic coherence related to several topics, including policy, in synchronous chatting systems. As our research involved a one-shot experiment,

[1] Throughout this chapter and the next, I use the first person plural pronouns "we" and "our" to refer to me and my research collaborators.

participants did not actually engage with each other in discussion. However, based on existing research about the content of comments that tend to receive a response in online discussion groups, we expect that topically coherent contributions would likely elicit participant interaction [3, 104].

While the case study focuses on just one deliberative concept, writing this article more deeply connected me to the considerations involved with applying deliberative concepts to an analysis, in general. By working through the concepts in this article with my collaborators, I also realized the possibility of designing to promote specific deliberative concepts, as hinted at in the discussion: "*One might curate (or moderate) for many of these values, including civility and quality [214], soliciting both objective and subjective descriptions [141], supporting both social and task processes [76], eliciting both logical arguments and situated experiences and stories [180, 185], and so on.*"

An important caveat is that *cognitive dissonance* [65] is used to describe the psychological mechanisms at play in this study, but in hindsight, this is not an appropriate framing. It is often difficult to make a decision when the options are close in their desirability. When pressed to do so, cognitive dissonance theory suggests that people are likely to experience a higher degree of dissonance than if the options are less close in their desirability. For example, a cappuccino and a latte may be more similar in their desirability than a cappuccino and a shot of printer ink. Choosing a cappuccino over printer ink is less likely to produce as strong a response to the rejected alternative, as the burdensome choice between a cappuccino and a latte.

As part of the protocol for this case study, participants were presented with two options about what should happen when a human-intelligence task (HIT) is rejected, either a *partial payment* or a *second chance* to complete the task. AMT crowd workers (called "Turkers") tend to prefer a second chance over partial payment, as observed in prior research [153, 152] and confirmed in the participant survey results. To consider how contributions to the topic coherence of an online policy discussion are affected by curating for positions in opposition to a participant's initial perspective, all participants were exposed to 20 seeded comments about the less desirable option—half in support and half in opposition to partial payment.

Several experimental design decisions make it difficult to read the study results as depicting a cognitive dissonance. First, it is less clear how salient the options were for participants. For example, we could have instructed our participants that we intend to reject their work on the task associated with the study, and then presented the options as a choice about how we handle their rejection—*would you prefer to be whipped or bludgeoned?* Second, we did not ask participants to rank their preferences, so it is also less clear how willing the participants who prefer second chance would be to accept partial payment, if offered the option over nothing at all. For this reason, it is tricky to interpret how desirable participant's viewed the pair of options. Third, we did not study contributions to a discussion seeded with comments about second chance. Controlled experiments are expensive, and for all the reasons discussed in the case (sections 5.3-5.5), curating for opposition was the more interesting scenario.

Specifically, it is less clear whether the results reflect a dissonance related to the pair of options (i.e., second chance, partial payment) or different interpretations of the partial payment policy topic itself, as some Turkers view the option as some compensation for their time and effort on a rejected HIT, but other Turkers view the option as an existential threat to the AMT labor market. Despite this limitation, the case does draw attention to factors beyond position agreement that affect new contributions to the topic coherence of an online policy discussion, which is useful to consider when evaluating the curation options at a discussion system.

5.2 Introduction

Coherence in online policy discussion refers to the *consistency of the topics within a thread of comments* [208]. When discussion participants are regularly off-topic, or move away from topics too quickly, the behavior leads to an incoherent discussion that cannot deeply consider a policy issue. Too much attention on a single topic is also limiting, although in practice online policy discussions often result in far more than one topic being deeply considered [42, 48, 84, 94, 135, 222].

Low coherence in online discussions about political issues has been observed across various forms of digital media: newspaper comment threads [48, 135], political [42] and non-political discussion forums [84], social platforms [94, 140], and field trials of advanced policy deliberation systems [86]. Although there are many recommendations [11, 40, 43, 219, 225] and prototype designs to encourage people using an online discussion system to build on the existing discussion [115, 119, 230], this empirical research indicates that *supporting coherence remains an unresolved design challenge*.

Here we examine the relationship between coherence and a commonly referenced design lever to affect the ways that people contribute to an online discussion: *content curation* [5]. Curating a comment thread means choosing which comments to include [48, 225], how to order those comments [124], and when to present them [153]. Such curation is useful when the chronological order of a comment is less relevant than its content (e.g., demoting profanity [48, 124] or highlighting political opinions [167, 166]).

Much of the work on comment curation has focused on how a reader's agreement or disagreement with the content presented affects their willingness to read or engage with it [171, 181, 229]. This is a tough sell, as most people prefer agreeable content most of the time [129, 166]. When people encounter content that challenges their own position, people may downvote it [33, 124], request fact-checks [120], or actively avoid it [59, 65]. This tendency to avoid disagreeable content raises questions about whether, when people do participate in discussions that feature positions contrary to their own, they are more likely to go off topic or introduce new topics, i.e., be less coherent.

In this paper, we extend this line of work to explore how curating comments around particular positions on a policy issue might affect new posters' willingness not just to reply but to reply coherently, increasing the chance that their comment furthers the discussion. We present results from an experiment that asked Amazon Mechanical Turk (AMT) crowd workers ("Turkers") to consider a proposed policy amendment to the AMT participation agreement to offer *partial payment* for rejected work. This proposal was presented in the context of an online discussion where comments were curated to prioritize arguments *for* or *against* the policy. We collected participants' initial positions of *support* or *opposition* to the policy as well as their comments and examined how agreement with the position in the curated comments correlated with their own comments' coherence with existing topics.

Overall, participants were less likely to add comments that cohere with existing discussion topics when the thread curation disagreed with their initial perspective. However, this was largely driven by people who disagreed with the proposed partial payment policy, who were especially unlikely to contribute comments that cohere with existing topics when seeing a thread that prioritized support for partial payment. By contrast, people who agreed with the proposed partial payment policy were more likely to add topic-coherent comments regardless of whether the curated comments were for or against partial payment. This *asymmetric* relationship between comment curation and coherence with opposition in a policy discussion suggests that designers of both discussions and discussion forums need to consider factors beyond whether a person agrees with a particular position when considering how to support effective participation.

5.3 Coherence and Curation in Policy Discussions

In a discussion, topics advance as participants reply and respond to each other along a common thread of subjects [13, 81, 208]. In this context of analysis, coherence is a function of how recent comments remain on the same topics introduced by the existing comments, which "seed" the discussion [208]. Without a coherent discussion of the pros and cons of a policy topic, it is impossible for a deliberating group to carefully weigh a policy issue [21, 76, 207].

Policy deliberation scholars have developed a few research methods for studying how groups of people talk with each other during a discussion about policy or civic issues [13]. For example, the Discourse Quality Index (DQI) is a communication coding scheme that is used to understand a policy discussion in terms of the *speeches* that people make during a discussion and how others might respond (e.g., with interruptions, counter-argument, or incivility) [203]. While the DQI is useful for studying the range of positions and level of respect during a policy discussion, a single speech may incorporate multiple topics to present a cohesive argument.

As an alternative, policy deliberation scholar Stromer-Galley [207, pg. 9] has developed a communication coding scheme to study policy discourse at the *thought* level of analysis: *A thought is defined as an utterance (from a single sentence to multiple sentences) that expresses an idea on a topic. A change in topic signaled a change in thought.* Stromer-Galley and Martinson [208] expand on the definition of topic, to characterize thoughts that add new topics to the discussion, versus thoughts that address the materials that establish the policy issue discussion (called "structuring topics") or thoughts that address topics that emerge through the ongoing exchange (called "interactional topics"). Stromer-Galley and Martinson [208, pg. 201-205] apply what they refer to as a *dynamic topic analysis* to measure *coherence* with the interactional topics, by tracking whether new thoughts add to or divert from topics already seeded in the discussion.

We chose these methods of characterizing and measuring coherence for multiple reasons. First, Stromer-Galley [207] has been applied in research revealing a lack of coherence in online discussion forums [84, 94]. Second, we found that the analytic granularity of distinguishing between coherence with the structural versus interactional topics was useful in our experimental design, which controls both the structuring and available interaction topics in the seeded discussion thread. Third, when people do post their thoughts to an online discussion forum, they are often in the form of comment(s), which felt closer to Stromer-Galley's definition of a thought than the DQI's notion of delivering a speech.

Coherence is one of a much broader set of concerns around deliberative discussion [13]. In this paper, we zoom in on coherence with the interactional topics because *talking-with*, and not *-past*, others in discussion is a fundamental precursor to deliberation [21, 76, 108] that is rare in online discussions [42, 197, 226], even more so when people disagree [129].

5.4 Curation and Disagreement

Disagreement is useful for small group and public decision-making. To quote John Stuart Mill's argument for why groups should not ignore opposition, "*If the opinion is right, they are deprived of the opportunity of exchanging error for truth; if wrong, they lose what is almost as great a benefit, the clearer perception and livelier impression of truth produced by its collision with error*" [157]. However, the individual experience of disagreement in a group can lead to feelings of threat [64, 129, 192], and people's reactions to these feelings can negatively affect the group [161, 127, 216]. This tension between value to group and threat to individual arises in a number of computer supported cooperative work (CSCW) contexts where people conflict with each other [58].

One potential lever designers have to manage disagreement and encourage engagement is comment curation: they can choose which comments are displayed, when, and in what order. Common strategies include showing the most recent comments, the most popular comments, and recommending personalized comments people would prefer to read [5, 166]. Many of these strategies, notably the popularity-based and personalized algorithms, tend to give people more of what they already like [176].

In this article we examine how coherence in an online policy discussion is affected by curating the discussion to promote either pro or con statements about a policy [124, 166]. Thread curation is particularly important when there are many [124] (and redundant [115]) comments in the discussion. In a policy context, thread curation can also be applied as a civic engagement lever to expose people to different views of an issue [129, 118, 167, 219, 197].

However, there is a tension in just how much opposition to present [166] and how its presentation affects a person's willingness to express their view [171, 181, 229]. This is especially risky for curation strategies that favor one side of a position over another, as might happen when trying to choose comments based on agreement or disagreement with a given participant's position, or to ensure that a particular view is heard. Further, even position-neutral strategies such as chronological order might naturally lead to situations where the discussion appears to be tilted toward one side or another, simply because the most recent subset of comments tend to agree [124].

5.5 Curation and Cognitive Dissonance

Managing the amount of disagreement present in a curated comment thread is important because it is easy for people to avoid disagreement online [129]. While online discussion can provide people with an opportunity to form community around shared values [50, 92], properties of digital environments also enable people to stay silent among the "invisible audience" of a policy discussion [16, 72, 184].

There are various reasons why people remain silent in the face of opposition. When a person's views are challenged they can experience *cognitive dissonance*, which can be unsettling and elicit an avoidance response [59, 65, 170], as people generally prefer not to be challenged [166]. Many people also feel unable to argue their positions, either due to a lack of training in argumentation, lack of leisure to study a particular policy matter [108, 192], or social risks of stating a position publicly [172, 197, 216].

These factors can discourage engagement with discussion topics when people see opposition in the thread, but might also encourage alternate forms of engagement that actively avoid or reduce the dissonance, such as by up/down voting comments [33, 124] or issuing fact-check requests [120]. Coe et al. discuss how people use such lightweight discussion system features in place of explicit disagreement: *"[...] users often used this thumbs up/down metric in place of expressing explicit agreement/disagreement within the text of a comment."* [33, pg. 676]

This leaves open the question of when people do add a comment to a discussion thread that prioritizes views counter to their own, whether their tendency toward avoidance translates to responses that are less coherent with the existing discussion. Drawing on cognitive dissonance theory, we argue that participating in a comment thread prioritizing "agreeable" content [166] will be more pleasant [59, 65] than one that presents disagreement, and that people will be more willing and able to coherently engage an existing discussion that is on comfortable, familiar ground. Further, we would expect people who disagree to tend to change the topic in order to reduce conflict between the expressed positions and their own.

Hypothesis: *Contributions to a policy discussion are more likely to cohere when participants are exposed to a thread that prioritizes comments that match their initial position.*

5.6 Method

We examine the relationship between comment curation, level of agreement, and coherence in an online policy discussion via an experiment with Turkers participating in a policy discussion about the AMT participation agreement.

5.6.1 Interface Design

We used a discussion forum interface modeled after RegulationRoom, a platform for civic engagement in public policy-making [185]. The interface included two panels: a summary of the AMT Participation Agreement and proposals to amend it on the left, and the comment discussion thread with a comment box on the right (Figure 5.1). Like RegulationRoom, the interface did not include *up/down* voting or other lightweight mechanisms for engaging with the content as our focus was specifically on topical coherence of comments rather than other behaviors.

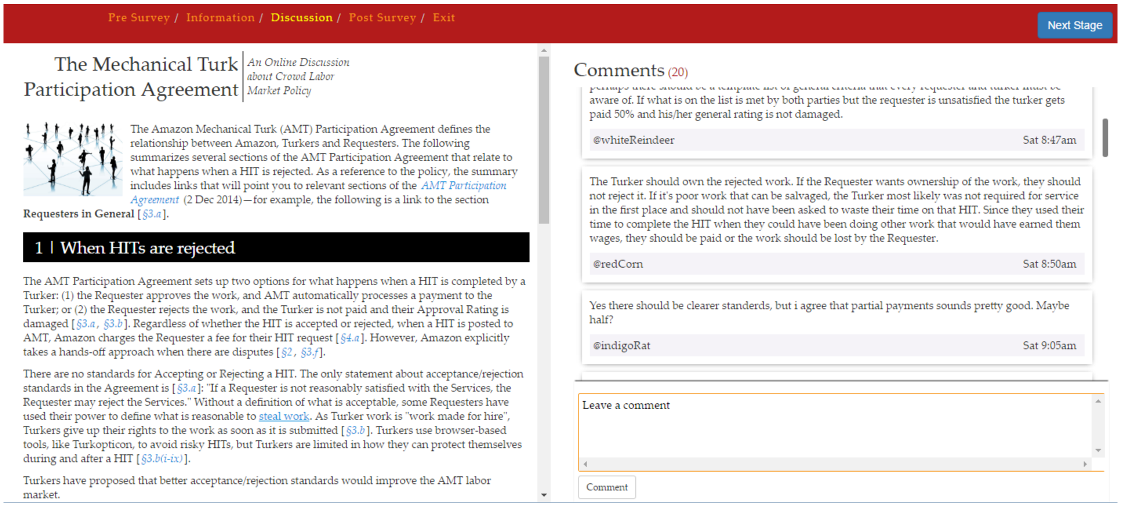
To set basic interface design elements we first prototyped the interface. We varied the placement of the comment text box (above or below the discussion thread) and the length of the discussion thread (short, with 3 seeded comments, or long, with 20 seeded comments that required scrolling) and tested these design variations in a pilot HIT.

A total of 408 Turkers accepted the pilot HIT, of whom 292 completed it (72%). Participants averaged 35 years old and about half identified as female; about 80% were U.S.-based. Participants were randomly assigned to one of the four conditions varying the comment box position and the discussion thread length. Participants were more likely to enter a comment when exposed to the longer discussion thread (OR 9.914, *p* ¡ 0.001). We found no effect of the comment box position on the likelihood to enter a comment. We decided to place the comment text box below the thread (as in Figure 5.1) based on eye-tracking research about how people read and skim articles online [56] with the hope that it would increase the likelihood of reading and engaging with other comments.

5.6.2 Materials

We developed the policy information presented in the discussion interface based on a summary of the AMT Participation Agreement around rejected "Human Intelligence Tasks" (HITs) posted by Requesters. A key concern for Turkers is whether a Requester accepts their work on a HIT [152], as the AMT Participation Agreement grants this power to Requesters with no recourse for workers.

We chose this specific policy topic because it has a direct impact on Turkers' everyday lives [147, 103, 114, 152] and therefore increases the ecological validity of the study. Based on suggestions by Turkers in a prior study [152], we proposed two changes to the policy. In the first, *partial payment*, Turkers would be paid for parts of the work that were considered acceptable by the Requester. In



The Mechanical Turk Participation Agreement | *An Online Discussion about Crowd Labor Market Policy*

The Amazon Mechanical Turk (AMT) Participation Agreement defines the relationship between Amazon, Turkers and Requesters. The following summarizes several sections of the AMT Participation Agreement that relate to what happens when a HIT is rejected. As a reference to the policy, the summary includes links that will point you to relevant sections of the *AMT Participation Agreement* (2 Dec 2014)—for example, the following is a link to the section **Requesters in General** [*§3.a*].

1 | When HITs are rejected

The AMT Participation Agreement sets up two options for what happens when a HIT is completed by a Turker: (1) the Requester approves the work, and AMT automatically processes a payment to the Turker; or (2) the Requester rejects the work, and the Turker is not paid and their Approval Rating is damaged [*§3.a*, *§3.b*]. Regardless of whether the HIT is accepted or rejected, when a HIT is posted to AMT, Amazon charges the Requester a fee for their HIT request [*§4.a*]. However, Amazon explicitly takes a hands-off approach when there are disputes [*§2*, *§3.f*].

There are no standards for Accepting or Rejecting a HIT. The only statement about acceptance/rejection standards in the Agreement is [*§3.a*]: "If a Requester is not reasonably satisfied with the Services, the Requester may reject the Services." Without a definition of what is acceptable, some Requesters have used their power to define what is reasonable to steal work. As Turker work is "work made for hire", Turkers give up their rights to the work as soon as it is submitted [*§3.b*]. Turkers use browser-based tools, like Turkopticon, to avoid risky HITs, but Turkers are limited in how they can protect themselves during and after a HIT [*§3.b(i-ix)*].

Turkers have proposed that better acceptance/rejection standards would improve the AMT labor market.

Comments (20)

perhaps there should be a template list of general criteria that every requester and turker must be aware of. If what is on the list is met by both parties but the requester is unsatisfied the turker gets paid 50% and his/her general rating is not damaged.

@whiteReindeer Sat 8:47am

The Turker should own the rejected work. If the Requester wants ownership of the work, they should not reject it. If it's poor work that can be salvaged, the Turker most likely was not required for service in the first place and should not have been asked to waste their time on that HIT. Since they used their time to complete the HIT when they could have been doing other work that would have earned them wages, they should be paid or the work should be lost by the Requester.

@redCorn Sat 8:50am

Yes there should be clearer standerds, but i agree that partial payments sounds pretty good. Maybe half?

@indigoRat Sat 9:05am

Leave a comment

Figure 5.1: The experiment interface, showing the policy summary on the left and the discussion thread on the right. The three comments that are initially visible and closest to the textbox are controlled to be either pro- or anti- the partial payment proposal.

the second, *second chance*, Turkers would have the opportunity to fix their errors in a rejected HIT.

We constructed the experiment so that the policy summary material (left pane of Figure 5.1) presented both partial payment and second chance; however, the comments that were seeded into the discussion thread (right pane) were exclusively about partial payment. We did this because there was more disagreement about the partial payment proposal, and because by focusing the discussion topics on partial payment, we were able to easily identify when new topics or structuring topics, like second chance, were introduced to the discussion.

To populate the discussion thread, we selected 20 comments contributed in the prior study [153], half in favor of and half opposed to partial payment. We then chose three comments pro- and three anti-partial payment as the focus comments that would be initially visible, according to the experimental condition. We added timestamps to make the discussion look recent and assigned pseudonyms (a concatenated color and animal, e.g., @blueMonkey) to each seed comment.

5.6.3 Participants and Recruitment

The HIT description recruited Turkers to test the user interface of a new online discussion forum platform. To attract viewpoints from a broad audience of Turkers, we did not restrict access to the HIT (e.g., to Turkers from specific countries or with specific levels of experience).

A total of 201 Turkers accepted the HIT, with 147 completing it (73%). On average, participants were 36 years old, about half identified as female, and 77% were U.S.-based. Turkers were paid $3 for their participation; the average time to completion was 17 minutes, resulting in a pay rate of about $10 per hour, a bit above the local state minimum wage. This payment structure adheres to the *WeAreDynamo* guidelines for Fair Payment in Academic Research[2].

5.6.4 Procedure

Upon accepting the HIT, participants were presented with a pre-survey that asked about their Turking experience, variables we used as controls in our quantitative analyses. They were also directed to select a pseudonym similar to those in the seed comments, using a random name generator that concatenated colors

[2] http $//wiki.wearedynamo.org/index.php/Guidelines\_for\_Academic\_Requesters$

with animal names, e.g., *@blueMonkey*.

Prior to entering the discussion interface, participants were informed that the discussion would be "about what happens when a HIT is rejected" and that as part of the experience they will "have an opportunity to take part in the discussion." We informed participants that the intent is to help resolve a lack of consensus among Turkers around the proposals that was observed in prior research [152]. They were asked to rate their initial position toward the two policy proposals on separate 5-item scales, from strongly disagree to strongly agree.

Participants were then placed in the discussion forum. They were able (but not required) to read a summary presentation of the relevant part of the AMT Participation Agreement and a description of the policy options, read a set of comments seeded in the simulated discussion, and add comments of their own. Participants were required to spend a minimum of one minute in the experiment interface; the average dwell time was 4.9 minutes (SD 3.6).

5.6.5 Curation Conditions

Each participant was randomly assigned to one of the following two conditions, each presenting the same twenty seeded comments, but sorted so that the first three comments emphasized different views toward *partial payment* (PP).

- *Pro-PP*: Three seed comments ordered closest to the comment text box presented support for partial payment.
- *Anti-PP*: Three seed comments ordered closest to the comment text box presented opposition to partial payment.

The specific comments for each condition are presented in Tables 5.1-5.2. We randomized the order of the three comments to control for order effects and separately randomized the order of the other seventeen comments. We realize that binary categorizations as pro and anti (or agreement and disagreement) are simplifications, and that real policy discussion and positions are often more complicated. However, most prior research and many real discussion contexts do have this binary flavor, so we adopt it as well; we will return to it in the discussion.

The curated comments for each condition were selected because they share not only a similar position (Pro- or Anti-PP), but to the extent possible, similar topics in the discussion. The Pro-PP comments relate to the Partial Payment Amount (Topic 4 in our analysis; see *Coding for Topic Coherence*). The Anti-PP comments relate to the Hands-Off Labor Market (Topic 5).

While the comments are different in their position on partial payment and topic, we did not control for other characteristics of the comments (e.g., character length, expressiveness). Without an *a priori* argument about how positions on questions about the AMT participation agreement would affect Turker responses, we chose to expose Turker participants to these positions in the unaltered words of other Turkers.

5.6.6 Ethical considerations

Tasking AMT Turkers to discuss topics related to the AMT Participation Agreement is a familiar research context for studying systems that support policy engagement and deliberation [118, 119, 153, 191]. While this context is conve-

Pro-Partial Payment Condition
#995: "I mean, for work of creative nature, a base pay should be fixed. If the requester keeps and uses the work, he should pay more."
#1094: "perhaps there should be a template list of general criteria that every requester and turker must be aware of. If what is on the list is met by both parties but the requester is unsatisfied the turker gets paid 50% and his/her general rating is not damaged."
#1136: "I believe that Turkers should receive atleast 25% of the task (if less than $5.00) or 10% (if more than $5.00) if it is rejected. However, they would need to have atleast shown effort and not just sped through the task. I've spent quite some time on a few tasks only to be rejected for something that was not clearly stated in the rules or was completely false. I believe their should atleast be an appeal system."

Table 5.1: Comments selected to emphasize **Pro-PP** views toward **partial payment** (PP) in the thread curation experimental conditions.

Anti-Partial Payment Condition
#976: "No. Turker won't get any partial payment. If he completes the hit with prescribed instructions ,then he will get full pay otherwise rejection."
#1119: "Allowing partial payment is a slippery slope, since some requesters would simply reject and give partial pay to almost everyone, citing the quality of their responses or whatever. What we need is real moderation from Amazon when there's real abuse of the system, instead of telling us it's between us and the requester and not their problem."
#1342: "There should absolutely be clearer standards for rejecting hits and those standards should be put forth to the worker up front. Workers should be able to discuss why the hit was rejected and also able to make a case for any problem or mistake made. Unless the requester can prove that a worker was clearly just hurrying through I think a rejected work should be paid in full. If we start accepting partial payments for rejected work it will lead to requesters looking for anything to reject and then paying less than they had advertised. It could be a sticky downward spiral."

Table 5.2: Comments selected to emphasize **Anti-PP** views toward **partial payment** (PP) in the thread curation experimental conditions.

nient, it requires Turkers to respond to their unequal economic position in the AMT labor market [103, 114, 152]. Unlike traditional labor markets, it is not clear how to address crowd work labor disputes through existing regulatory authorities [64], and unlike other social platforms (e.g., Reddit [24]), Turkers are not well positioned to effect change in AMT [191]. Performing research in this experimental context therefore has special ethical circumstances that need to be considered.

We received IRB-approved informed consent from all participants and compensated their time based on Turker approved standards for academic research [191]. We also implemented several Turker-supported best practices for HIT design [151], as it is important to remember that Turkers participate as part of a *task* they perform for a *reward*. In addition to taking these measures to treat participants *fairly*, we also worked with the community manager at *TurkerNation* [147] to develop the specific policy language for the study to make the content of the policy proposals relevant and engaging to the participating Turkers. Finally, we indicated that our research group is not associated with Amazon and that the purpose of the experiment was purely for research.

5.7 Data Analysis

5.7.1 Coding for Topic Coherence

The twenty seed comments were chosen to cover six topics identified based on 1092 Turker comments from a prior study [152] using an affinity diagramming analysis process:

1. *HIT Design*: Unclear instructions or acceptance standards and technical errors should result in partial payment
2. *Requester Communication*: Lack of Requester-to-Turker communication
3. *Turker Quality*: Low quality Turker work should not be paid (e.g., completed too quickly, robot accounts)
4. *Partial Payment Amount*: Proposes an amount or scheme for implementing partial payment (e.g., 10%, 25%, 50%)
5. *Hands-Off Labor Market*: Amazon's "hands-off" approach to the labor market (e.g., partial payment could lead to more rejections or low quality work)
6. *HIT Specific Policies*: Different protocols for different tasks (e.g., Turkers should own or receive a base payment for rejected creative work)

To identify when a comment made by a participant cohered with topics in the discussion, we used a coding scheme based on Stromer-Galley's definition of topic coherence [207, 208]. Two coders independently categorized each comment as either "new topic" or assigned a set of Topic ID numbers (1-6) identifying the topics referenced by a comment. The two coders trained initially with a set of 95 comments, resolving disagreements during the training period. Training continued until the Cohen's Kappa score for inter-rater reliability was above 0.8 and then the coding was tested on a holdout set of 95 comments. The final Cohen's Kappa score was 0.85.

The following is a sample participant comment from the current study that coheres with Topic 4 (Partial Payment Amount): *"I think partial payment should be more like 85% rather than 10%. If you only get 10% for partial payment, then I'd* *probably rather just redo the HIT."* [P69][3]

As an example of a comment that *did not* cohere with the seeded comments: *"I liked the idea of a second chance better than partial payment. I would like the chance to fix my mistake (if I make one). I'm honest. When I answer surveys, I read every question. I don't randomly just choose answers."* [P14] This comment does not address the seeded topics, as it raises the *second chance* proposal which was excluded from the discussion thread. The response is somewhat related to Topic 3 (Turker Quality), though it does not speak to the specific concern that offering partial payment encourages low quality work. Unlike the prior example, the response does not provide any contextual markers that connect it to any existing interaction topic, such as the brief comparison of *10%* vs. *85%* that indicates that the prior comment coheres with Topic 4.

5.7.2 Metrics

Response Variable

The response variable used the above coding scheme to examine if a comment coheres with topics in the discussion or not.

- *Coherence (relating to existing topics)*: A hand-coded binary variable at the comment level capturing whether the comment coheres to one of the six topics addressed by the existing seeded comments.

[3] Comments are associated with a unique identifier of the participant ranging from P1 to P147.

Independent Variables

Independent variables were based on the experimental conditions and initial survey responses to the partial payment proposal.

- *Curation Condition (Pro-PP, Anti-PP)*: Captures whether participants were exposed to a discussion prioritizing comments that were pro- or anti-partial payment.
- *Initial Position (Support, Neutral, Oppose)*: Participants who rated their position toward partial payment as strongly agree or agree were coded as *support*; those who rated as strongly disagree or disagree were coded as *oppose*; others were coded as *neutral*.

For modeling "simple agreement" (see Table 5.5) we combine the Curation Condition with Initial Position into a single Matching Preference variable.

- *Matching Preference (True, False)*: Captures whether the participant's Initial Position matched the Curation Condition (i.e., Support x Pro-PP or Opposed x Anti-PP). For this analysis of simple agreement, we removed participants with a "Neutral" view.

Control Variables

At the participant level, we controlled for participants' self-efficacy, geographic location, and their past experience with rejections.

- *Self-efficacy*: Eight scale items of generalized self-efficacy and confidence in one's own abilities and skills [26] were averaged into a single measure (Cronbach's α = 0.93). In prior research, newcomers to an online policy discussion with high assessments of their own self-efficacy contributed comments that were longer and more responsive to the policy topics [153].[4]
- *Country*: A binary variable coded as 1 for United States-based participants and 0 for others.
- *Rejected HITs*: In the pre-discussion survey, participants estimated the total number of HITs they have had rejected. We centered and standardized this variable, such that a one unit increase in Rejected HITs reflects a one standard deviation increase in the variable.

At the comment level, we controlled for comment length.

- *Total Words*: Total number of words in a comment, centered and standardized in the same way as Rejected HITs.

5.7.3 Statistical Models

As some participants made multiple comments, we treated *participant* as a mixed-effects nesting variable to account for non-independence. Mixed-effects logistic regressions were used to predict topic coherence at the comment level,

[4] We chose to use generalized self-efficacy as opposed to a context specific self-efficacy measure because research shows feelings of ability can translate across contexts [7]. Further, our task required multiple specific efficacy constructs (e.g., reading efficacy, writing efficacy, political efficacy); if we were to choose one, it is unclear which would be the most appropriate to measure, and measuring several would introduce extra burden on participants.

as a binomial distribution was appropriate for the binary response variable. Model-level significance was evaluated using the log-likelihood ratio test, the Akaike Information Criterion (AIC), and the Bayesian Information Criterion (BIC).

The model coefficients are interpreted as the expected change that each independent variable contributes to the logits of the response variable. In the findings we exponentiate the logits to present the odds ratios. Odds ratios can be interpreted as the change in the response variable expected from a one-unit increase to an independent variable, holding others constant.

However, when evaluating the effect of an interaction, the coefficient estimated for the interaction is added to the main effect of the interacted variable. The combined effect of the interaction can then be exponentiated to present the effect as an odds ratio, i.e., exp(*main effect* + *interacted effect*). After a model is fit to the data, the model can be used to estimate the expected likelihood of the dependent variable at various levels and combinations of the coefficients—these expected values are in terms of predicted marginal means. We use a Tukey-based pairwise comparison of the expected marginal means to examine the interactions within the models (using a 95% confidence interval).

5.8 Findings

5.8.1 Descriptive Overview

Table 5.3 presents the descriptive statistics for the study. As part of our question is whether initial position might influence behavior, we first confirm that the distribution of initial positions was not significantly different between the curation conditions: $\chi^2(2, 147) = 3.013$, $p = 0.2217$.

Table 5.4 reports the number of participants, comments, and coherent comments by condition. Although the task instructions explicitly did not require participants to leave a comment, most participants did so, with just under half cohering with existing topics (46.4%). We found no significant difference in the likelihood to make a comment by condition: $\chi^2(1, 147) = 1.9e\text{-}29$, $p = 1$. Therefore, we focus on the likelihood to cohere with the existing topics within the discussion thread.

5.8.2 Agreeing with Curated Position Increased Coherence

The data support our hypothesis that presenting participants with content that matches their initial position increases the likelihood of contributions that cohere with the existing discussion topics. Comments made by participants whose initial position matched the curation condition were 2.970 times more likely to cohere with the existing discussion ($p < 0.05$, see Table 5.5). This mirrors HCI research about recommending content that is agreeable [166] and similar to what a user already likes [5, 176].

Participation	*Count*
Accepted HITs	201
Completed HITs	147
Commented	139
Total Comments	155
Control Variables	
Self-Efficacy (*Mean, SD*)	2.7 (0.46)
Country: US-based (*Count, Percent*)	122 (82.9%)
Rejected HITs (*Mean, SD*)	84.54 (258.43)
Total Words (*Mean, SD*)	296.8 (239.41)
Initial Positions for Partial Payment	*# Participants*
Support	53 (33.1%)
Neutral	39 (28.1%)
Oppose	55 (38.6%)

Table 5.3: Descriptive statistics capturing participation in the experiment, control characteristics, and details about participant initial position for partial payment prior to the discussion stage of the experiment.

	Pro-PP	*Anti-PP*	*Total*
Participants	74	73	147
Total Comments	81	74	155
Coherent	38	34	72

Table 5.4: Descriptive statistics capturing the count of participants, comments, and coherent comments in each curation condition (Pro-PP, Anti-PP). Almost 95% of participants contributed a comment.

5.8.3 Opposed Positions were Overall Less Coherent

However, the significant *Intercept* in Table 5.5 and relatively weak significance of the *Matching Preference* covariate indicate that the model is missing a good deal of variance. Thus, we next examine a model that distinguishes between participants' initial positions (Table 5.6). This analysis shows that beyond agreement or disagreement, initial position matters. Participants *Opposed* to partial payment were significantly less likely to post responses that cohere with the discussion topics than those who *Support* it (OR = 0.173, $p < 0.01$). Further, *Opposed* participants in the Anti-PP condition, who see comments they agree with, were significantly more likely to cohere than in Pro-PP: exp(-1.753 + 2.224) = 1.600 OR ($p < 0.01$). This finding aligns with the primary argument, that agreement and coherence go hand in hand, although initial position also helps predict coherence.

5.8.4 Effects of Initial Position were Asymmetrical

Because interaction effects where there are multiple levels can be tricky to evaluate, we next applied a Tukey-based pairwise comparison of each of the variable levels (e.g., comparing coherence likelihood between initial Neutral and Support positions, while keeping the curation condition fixed at Anti-PP). Figure 5.2 graphically depicts the analysis as a predicted marginal means interaction plot between discussion curation (Pro-PP and Anti-PP) and initial position for partial payment (Support, Neutral, Oppose).

Two main points emerge out of this analysis. First, the Pro-PP curation condition generated significantly more *coherent* contributions from those with *Neutral* or *Support* initial positions compared to those *Opposed*, 7.054 times ($p < 0.01$) and 5.292 times, ($p < 0.05$) respectively, while the Anti-PP curation condition did not show this difference. Second, the main driver of this effect is those with an *Opposed* initial position, as their contributions were significantly less likely to cohere, by 0.228 times ($p < 0.01$) in the Pro-PP condition versus Anti-PP. This difference between the conditions does not occur in other cases.

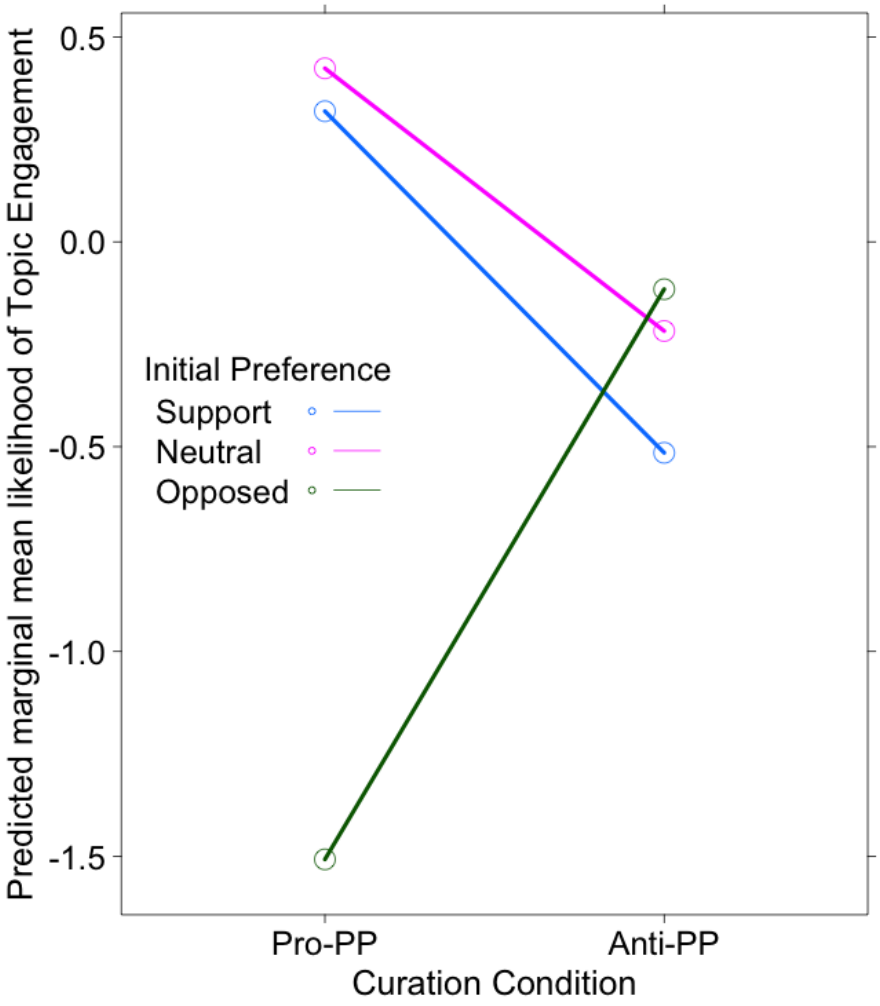


Figure 5.2: Predicted marginal means interaction plot between discussion curation (Pro-PP and Anti-PP) and initial position for partial payment (Support, Neutral, Oppose). Linear predictions are given on a log scale and reflect the estimated probability of a topic engagement.

	Coherence	
	Est (SE)	*OR*
Curation Condition and Initial Positions		
(Intercept)	-2.691 (1.54)	0.067 .
Matching Preference	1.088 (0.44)	2.970 *
Control characteristics		
Self-Efficacy	0.691 (0.54)	1.995
Country: International	-0.338 (0.57)	0.713
Rejected HITs	-0.328 (0.71)	0.720
Total Words	0.408 (0.18)	1.503 *
Log likelihood	-70.02 (df=7)	

Table 5.5: Fixed-effects logistic regression predicting the likelihood that participants will engage specific seeded discussion topics when their expressed preference matches the curation condition (i.e., *Pro-PP x Support*, *Anti-PP x Opposed*). *p-value significance*: *** 0.001; ** 0.01; * 0.05; . 0.1;

5.9 Discussion, Implications, and Limitations

At a high level, the findings confirm our hypotheses that curating a discussion thread to match a participant's preferences increases the likelihood of contributions that cohere with the existing discussion. This observation extends existing research about how people prefer content that agrees with their preferences [129, 166, 176], by demonstrating how curation can affect whether new contributions cohere with or diverge from the existing discussion. However, the effect of comment curation and participant preference was asymmetrical: people who support partial payment were much more likely to engage coherently with the conversation when they saw comments curated to present the Anti-PP position, compared to people opposed to partial payment who saw the Pro-PP curation condition.

This observation implies that curating primarily for position consistency led us (and presumably, will lead others) to downplay other aspects of comments

	Coherence	
	Est (SE)	*OR*
Curation Condition		
(Intercept)	-1.215 (1.26)	0.296
Anti-PP	-0.850 (0.60)	0.427
Initial Position for Partial Payment		
Neutral	0.208 (0.65)	1.232
Opposed	-1.753 (0.64)	0.173 **
Control characteristics		
Self-Efficacy	0.615 (0.42)	1.850
Country: International	-0.597 (0.53)	0.550
Rejected HITs	-0.568 (0.76)	0.566
Total Words	0.443 (0.17)	1.557 **
Interaction (Curation Condition x Initial Position)		
Anti-PP x Neutral	0.043 (0.92)	1.286
Anti-PP x Opposed	2.224 (0.85)	1.600 **
Log likelihood	-94.41 (df=10)	

Table 5.6: Fixed-effects logistic regression predicting the likelihood that participant responses will cohere with specific seeded discussion topics. Initial Position terms are by comparison to *Support* for partial payment and Curation features are by comparison to the *Pro-PP* condition, both of which are baselines in the Intercept. The odds ratios reflect the exponentiation of the estimates for each feature. *p-value significance*: *** 0.001; ** 0.01; * 0.05; . 0.1;

and contributors that might affect people's coherent engagement with diverse perspectives on a policy issue. For instance, Munson and Resnick found that reaction to individual items in a curated list of news articles depended not only on that single item but also on the others surrounding it [166]. Moving towards implementation, the discussion platform *ConsiderIt* [118] allows an individual user to adopt a variety of arguments both for and against any given policy decision.

This brings back to our choice of binarizing positions and comments on the policy into a pro versus anti framing, as such work shows how, in practice, reactions rarely occur in response to a single, isolated policy proposal. (Indeed, for that reason we originally planned to examine both second chance and partial payment rather than focusing on just one proposal. However, that plan was ultimately discarded, both due to the complexity of the analysis and the limited comparability to prior work.)

Taken together, these results and work suggest that research and design around engaging with disagreement would benefit from more nuanced views and analyses of how and why people agree or disagree with the positions proposed. Tools such as ConsiderIt, or argument analysis techniques from the area of natural language processing, could lead in fruitful directions for better supporting contentious policy discussions.

5.9.1 Curating for openness rather than agreement?

Our results also raise an important caveat around comment curation: curating comments risks systematically excluding viewpoints. The general versions of this are familiar: popularity tends to curate for and perpetuate majority opinion; personalization tends to curate for agreement, which may increase engagement but primarily with like-minded people [176]. In our case, although people in all conditions were equally likely to contribute a comment, people who were Opposed to partial payment were much less likely to cohere with the current discussion (especially when presented with Pro-PP arguments)—and contributions seen as off-topic are often ignored.

In this experiment, one possible driver of the asymmetry was a difference in the specificity of the featured Pro-PP and Anti-PP comments. Pro-PP was more specific around implementation decisions—i.e., setting the right level of partial payment—while Anti-PP comments were more general, describing how the partial payment proposal would have damaging effects on workers by giving HIT Requesters more room to reject requests.

We posit a parallel to how high fidelity interface prototypes tend to elicit comments about specific design elements, versus napkin sketches that give more room for considering the overall interaction [22]. Proposals that focus attention on specific implementation details will tend to concentrate attention on those details, arguably leaving less room for coherent discussion on topics where there is disagreement than more open discussions of the policy context. Opponents seeing Pro-PP curation may also have felt like they were joining the planning committee for a distasteful proposal—an unlikely scenario for effective, coherent contributions—while supporters seeing Anti-PP curation may have felt like their opinion on the issue still contributed to its deliberation.

In an attempt to integrate topic coherence [208] with common definitions of deliberation, such as careful weighing of diverse perspectives [21, 76, 207], we suggest that discussion and deliberation moderators might want to encourage (support) coherent disagreement: *a thread of comments that consistently contribute to a careful weighing of differing perspectives on a topic*. Designing to support coherent disagreement might imply highlighting content that leaves more room for debate (just as good interviewers, and bad lawyers, ask open-ended questions). How well this strategy for promoting coherent disagreement would work is an open question—choosing a controlled, one-shot experiment on one topic means we cannot make strong claims to generality or ecological validity—as is the question of what properties of a comment would invite openness.

Still, the idea has potential and is worth further study, both at the level of individual comments and of groups of them. Online discussion moderation means managing a stream of comments, often one at a time [48, 123, 124], or engaging specific comments to learn more about the experience of specific commenters [185, 61]. Supporting coherent disagreement might mean curating (or moderating) sets of comments based on characteristics of the group: expressing a range of positions, possessing topical coherence as a group, affording overall openness to discussion, representing a diversity of stakeholders, and so on.

5.9.2 Changing curation strategies and metrics over time

Another consideration is that the goals of a deliberation change over time—and another reading of the curated comments in this study is that the Pro-PP condition presented the deliberation as farther along than the Anti-PP condition. As the state of a deliberation transitions from investigating a common problem, to eliciting a range of potential ideas, to critiquing the ideas and refining them into a single proposal, each transition starts to impose constraints on the discussion topics. Thus, the discussion naturally tends to narrow through proposal development, making it harder to make coherent contributions, especially for those who disagree with the fundamental approach rather than with some implementation detail.

The shifting nature of policy discussion and deliberation group tasks over time suggest that curation strategies and metrics should likely change with them. We focused on coherence in this paper because it is understudied [71] and because *talking-with* the existing discussion is a common value in many policy discussions [42]. However, our measure of coherence is only appropriate for some goals. In our definition, the opposite of being coherent was not "incoherent”, but could include introducing novel topics and ideas; such divergent thinking has real value at many stages of many group processes [51, 52].

More generally, work around curation—including this paper—has tended to focus on the specific problem of curation to support engagement with disagreement. This is an important problem to be sure, but is a small part of a much wider range of deliberation desiderata [21, 76, 108]. One might curate (or moderate) for many of these values, including civility and quality [214], soliciting both objective and subjective descriptions [141], supporting both social and task processes [76], eliciting both logical arguments and situated experiences and stories [180, 185], and so on.

Curating for disagreement might come at the expense of other goals such as civility, social affect, or solidarity; putting disagreement at the center leaves these other important concerns at the margin.

5.10 Conclusion

Online discussions about policy are shaped by the ways that contributors dynamically add to, expand on, or divert attention away from existing topics in the discussion. Here we operationalize this dynamic by measuring coherence with the interactional topics in an online discussion, and predict higher levels of coherence when participants are exposed to a comment thread that prioritizes positions of the policy that match their own preference. However, we observed an asymmetric relationship between preference and curation, which implies that curating primarily for or against consistency with a poster’s current position may overlook other important aspects of comments, including their openness to coherent deliberation, that might affect people's willingness to engage coherently with diverse perspectives on a policy issue.

5.11 Acknowledgement

We thank Daniel Wu for developing the infrastructure for this research and Sara Chau for her support through the analysis. Thanks also to Kristy Milland and the Cornell eRulemaking Initiative (CeRI) team for their support developing the policy context for the AMT Participation Agreement discussion. This work is supported by NSF HCC 1314778.

CHAPTER 6

CASE STUDY 2: NEWCOMER CRAFTED POLICY DISCUSSION PROMPTS

6.1 Forward

The chapter presents research published at the 2018 ACM International Conference on Computer Supported Cooperative Work (CSCW) under the following citation:

McInnis, B., Leshed, G. and Cosley, D., 2018. Crafting Policy Discussion Prompts as a Task for Newcomers. In the *Proceedings of the ACM on Human-Computer Interaction*, Volume 2 (CSCW), p.121.

As with Case Study 1, first person plural pronouns are used throughout to refer to the research team responsible for this work. The chapter presents a case that demonstrates how to develop and evaluate a system feature modeled after a deliberative concept, specifically *meta-talk about conflict*. This design process involved researching the operational definitions related to the concept and then borrowing ideas from system design (i.e., newcomer onboarding processes, crowd-writing systems) to implement the system featured in the chapter.

6.2 Introduction

Research about online policy discussion routinely finds that people tend to talk *past* rather than *with* each other about their different perspectives on an issue

[42]. One route toward more effective discussion is to include meta-talk in the discussion [207]. Taken from the policy deliberation literature, meta-talk refers to talk about the current state of a discussion, such as its tone [14, 33], opportunities for consensus [169, 204], or points of conflict [207]. In face-to-face policy deliberations, which often include a professional facilitator, this moderator will actively listen to the conversation for opportunities to insert meta-talk, often in the form of a discussion prompt related to the ongoing conversation [145, 160].

This strategy is hard to directly transfer to online policy discussions because the scale of both audience and discussion is far larger than individual moderators can manage [61, 124]. In this paper, we explore whether the work of crafting reflective meta-talk discussion prompts could be distributed to the community. More specifically, we focus on newcomers as a resource for crafting these discussion prompts.

Newcomers are often seen as a problem for groups to manage, with needs around socialization and mentoring [29, 66, 100] and often different perspectives than existing members [39]. Different perspectives, however, can be useful: as naïve outsiders to a group, they can raise questions or observations that other members may have forgotten or willingly ignore [4, 30, 29, 132]. Further, Kraut et al. [116] call on system designers to leverage these potential benefits as newcomers investigate an online community. As such, newcomers' fresh perspectives and minimal social constraints may make them well-suited for creating meta-talk-based prompts.

To this end, we designed a task that newcomers complete before they enter a discussion. In the task, they are asked to create a prompt for group discussion based on a pair of comments with different perspectives on a policy issue [177].

The design is inspired by existing human-computer interaction research about listening systems, notably *Reflect*, a micro-task workflow designed to interrupt the way that people reply to each other in an online discussion by segmenting the reply-and-response cycle into micro-tasks modeled after active listening principles [119]. It also draws on elements of text summarization in iterative crowd-writing systems, e.g., *Turkit* [134]. It goes beyond these, however, by asking participants not just to restate or summarize existing text, but instead to craft new questions to further an ongoing discussion.

A task that asks newcomers to listen to others' comments may also have beneficial effects on newcomers' later participation in the discussion. Kriplean et al. [119] offer an untested provocation that, "[...] listening interfaces help establish an empathetic normative environment," proposing that, "if the interface can encourage some users to listen, others may follow, helping to establish constructive communicative norms" [119, pg. 2]. Thus, we also ask how performing the task affects newcomers' comments in the policy discussion: are they more explicitly reasoned [207] or coherent with existing topics [154], as these measures indicate an effort to talk with, rather than past, others [94, 140, 189].

To address these questions, we conducted a controlled experiment in which Amazon Mechanical Turk (AMT) workers participated in a HIT that included a simulated online policy discussion about the AMT participation agreement. We briefed participants about the policy issue, surveyed them about their initial perspectives and relevant backgrounds, and exposed them to one of several versions of the onboarding task design that varied in terms of the degree of disagreement in the comments they were asked to design prompts for, whether they were creating a new prompt or improving an existing one, and the amount and structure of the task instructions. After completing the onboarding task, participants entered a simulated discussion forum where they could read other comments about the policy and post their own comments. Before exiting the HIT, participants were polled once more about their perspectives on the policy issue.

To evaluate the discussion prompts, we created measures inspired by the style of prompts that teachers develop for social studies lessons that involve classroom-based deliberation (see [97, 148, 177]). We found that participants were often able to produce acceptable discussion prompts, such that the prompt identifies commonalities and points of comparison between the two comments and asks a genuine question about the differences. They were better able to do this when given more instructions on how to write them and comment pairs with more disagreement; and in variations of the task where they were improving an existing prompt, when given more open-ended prompts to work on. However, completing the onboarding task had no effect on the reasoning or topicality of comments participants later made to the discussion forum. In fact, participants in task variations without instructions contributed longer forum comments than those who received instructions, raising questions about how instructions affected the way participants allocated their effort between the prompt creation and discussion commenting portions of the study.

Taken together, these results suggest that newcomers may indeed be able to create discussion prompts that community moderators might use to further a policy discussion. Giving clear instructions dramatically increased people's ability to effectively complete the task, adding weight to the importance of clear task design in the practice of crowd work, while the findings about the value of difference and open-endedness can inform the design of other crowd-based systems to support synthesis and reflection. Finally, our lack of evidence for beneficial effects of a meta-talk task on newcomers' later contributions to the discussion raises questions for future work that hopes to more effectively onboard newcomers to discussion communities.

6.3 Background

6.3.1 Meta-talk

Definition

A policy deliberation is a group-based discussion activity where a group is challenged to carefully weigh the diverse views of its members on a policy issue [21]. Policy deliberation scholars use a number of discourse analysis methods to study how people talk with each other during a deliberation [13]. Particularly relevant to our goal of leveraging newcomers' insights to pose questions for an existing deliberation are the problem-talk and meta-talk concepts from Stromer-Galley's method for *Measuring Deliberation's Content* [207]. This method defines an effective deliberation as incorporating six elements: reasoned opinion expression, sourcing, disagreement, equality, topic, and engagement. These elements are identified through discourse analysis that monitors the discussion for shifts in topic and for the presence of specific forms of talk: problem-talk, meta-talk, process-talk, and social-talk.

Whether phrased as a reasoned opinion or a rhetorical question, when people make statements that advance a claim, such as adding facts or arguments to the discussion, they are contributing *problem-talk* to the deliberation. By contrast, *meta-talk* is "talk about the talk" that clarifies and identifies potential consensus or conflicts related to the problem-talk, "[...] that attempts to step back and observe what the participant thinks has happened or is happening and why it's happening" [207, pg. 12].

As mentioned earlier, newcomers may be especially able to contribute certain kinds of meta-talk to a discussion. Because of their lack of investment in and history with a group, they may be more likely than existing members to provide new framings of existing arguments, raise issues a group has avoided but that need to be addressed, or ask genuine questions that bridge competing opinions, especially if they come in without strong opinions of their own. There are potential downsides to the idea of newcomers working to generate meta-talk as well—they might pose tired, old topics that are well-settled, or inadvertently trip over taboos or negative relationships—but we still see this as an idea worth pursuing.

Online Discussion System Examples

Other online discussion systems have explored the value of encouraging meta-talk, though usually among existing members. Several such systems have focused on negative interactions and disrespect between existing participants, as meta-talk is often observed calling out such behavior [33, 14, 158]. For instance, *Mediem* [169] is described as a "deep dialogue discussion forum" that provides several mechanisms for reflective interaction during a group discussion. Using a "Conversation Thermometer", participants rate and reflect on the quality of their discussion at specific moments.

In a more specific attempt to introduce meta-talk—and one that might be especially apropos newcomers—the *Reflect* platform was designed to scaffold an active listening exchange between speakers and listeners [119]. Reflect augments an existing discussion forum by incorporating a micro-discussion about a comment, primarily to clarify statements made by a speaker. In the Reflect workflow, a listener briefly summarizes "what they heard" in a comment, highlighting specific sentences and phrases. The speaker then responds to the summarized points by clarifying and responding to the listeners' interpretations. Although aimed more at supporting one-on-one interaction than the kinds of group processes meta-talk generally targets, Reflect was a real inspiration toward our task design, and we use a similar process as a baseline onboarding activity that emphasized listening to others' comments without the meta-talk aspects of our design.

6.3.2 Crafting a Policy Discussion Prompt

Having argued for the general idea of asking newcomers to contribute meta-talk to a discussion, we now turn to the more specific question of what sort of meta-talk to contribute. Meta-talk is often provided in formal policy deliberations by professional moderators [108]. Moore [160] observes that moderators often "[follow] from the front," prompting conversation with a *good question*, then getting out of the way, "follow[ing] the group as it unfolds its own discourse on the issue at hand" [160, pg. 4].

Several strands of research have looked at criteria and sub-components that make for effective meta-talk discussion prompts. For example, from the discourse analysis perspective, Schiffrin [193] operationalizes meta-talk as having a few common linguistic components: a *reference* (e.g., phrases, sentences) from the existing discourse, which are accompanied by *logical operators* that evaluate a reference or compare a set of references, as well as verbs that *request* an action related to the reference(s), such as clarify, tell, or argue.

There are close parallels between Schiffrin's formulation and aspects of teacher training in social studies [90, 97] aimed at the craft of creating policy deliberation prompts [177, 148]. Social studies teachers often use these prompts in their classroom practice to facilitate policy deliberations among students. Parker [177] recommends that teachers craft a policy deliberation prompt with a formula similar to Schiffrin's: (1) introduce a common problem and how it is personally relevant to the members of a discussion group, (2) logically compare a set of alternative solutions, and (3) request that the group make a decision. These are illustrated in the following example prompt, about whether teachers should reveal their own views on a policy to their students [177, pg. 14]:

> "(**1. Common problem**) You are learning a number of ways to identify controversies that are at the core of the topics you are going to teach and also a number of ways to help your students study those controversies. (**2. Alternative solutions**) Do you believe that teachers who engage students in the study of controversial issues should reveal their own positions on those issues? Or is it better for teachers to keep their opinions to themselves? (**3. Request**) I wonder if we can come to a consensus on this."

This prompt both illustrates the formula and calls out a larger issue that needs to be considered in crafting policy prompts and in moderating deliberations, around the risks of biasing the discussion through both a prompt's content and how it is presented. Rhetorical questions, for instance, are considered to be problem-talk that advances a position rather than genuine questions that characterize meta-talk [207, pg. 25], while subtleties of both the verbal and body language of a moderator posing the prompt can influence deliberation participants' behavior [202].

This line of work around crafting effective prompts to support meta-talk informed our task and experimental design. The elements of listening to the existing discourse, finding points of comparison and connection to call out, and asking genuine questions related to them are directly embedded in our coding scheme for evaluating the quality of prompts, the task instructions we developed, and some variations of the task interface.

6.3.3 Onboarding Newcomers to an Online Policy Discussion

As previously mentioned, during a face-to-face deliberation, a trained moderator will use discussion prompts to encourage the group toward deep consideration of a policy issue [145, 160]. In online discussions, a team of moderators might work together to facilitate large audiences, using specialized systems that help the team to promote dialogue among many participants [61, 86]. Whether in face-to-face policy deliberations [21] or online policy discussions [124], newcomers typically do not perform this type of facilitation.

In fact, there are often few expectations of newcomers to an online policy discussion. Existing members will greet newcomers [100], listening to their concerns and providing feedback [19, 20]. In online collaborative work, existing members will offer newcomers training and tasks [29, 66] to help them feel useful and invested through peripheral, but legitimate participation [93, 223]. These and other tactics are used to encourage newcomers to become regular members, as an online community needs new members to grow [116]. However, this work can tax existing members [66].

Rather than relying entirely on existing members, Kraut et al. [116] argue that the design of an online system should help people "[...] make a decision about joining and to respond to the common moves that newcomers use when forming impressions of the community" [116, pg. 3]. During what is referred to as the *Investigatory* phase of group socialization, a newcomer will collect information about the group to predict whether it will fit their needs [161, 162, 131]; such investigation in online discussions often involves reading a potentially large number of comments to make sense of the perspectives already in the discussion [183, 115, 230].

While investigating a group, a newcomer might raise questions that help existing members to recognize new ideas from older discussion points (called *Newcomer Innovation*) [30, 132]. We suggest that an onboarding activity of crafting a discussion prompt around existing comments might support newcomers' investigation while encouraging such innovation, explicitly providing occasion for both listening and for raising questions. The activity might also elicit feelings of investment, by tasking a newcomer with work that has value to the community.

6.3.4 Leveraging Ideas from Crowd-Writing

One issue with asking newcomers to craft discussion prompts as an onboarding task is that it is effortful: crafting a discussion prompt can take a moderator a substantial amount of time [177]. As newcomers are not likely to be willing to invest too much time or effort in a new group, we surveyed the crowd-writing literature to consider task designs that might ease this burden. Examples of crowd-writing systems we considered include *CrowdForge* [113], *Knowledge Accelerator* [91], *Mechanical Novel* [111], *Soylent* [9], *Storia* [110], and *TurkIt* [134]. These crowd-writing systems demonstrate how a complex writing task, such as crafting a policy discussion prompt [177], might be structured as a sequence of briefer tasks.

In particular, we focus on ideas from iterative writing workflows. Little et al. describe the process of iterative writing through crowd work as having three main general sub-tasks: writing, improving, and evaluating pieces of text [134]. Providing clear instructions and criteria for the write and improve tasks implicitly supports self-evaluation and generally improves performance [113], while mitigating risks of low-quality work or unfair rejection of work in micro-task markets [152]. Explicit support for helping people self-assess their work has also been shown to yield better quality writing [53]. Support for evaluating others' work can also help guide workers to make better suggestions for improvement or recognize that no improvement is necessary [91].

Another consideration is that workers performing improvement tasks are likely influenced by the text they are improving. A piece of text offered for improvement implicitly communicates information about both stylistic norms [91, 230] and the thinking of previous individual contributors. In a policy deliberation context, it is possible that the same kinds of biases that Parker [177] was concerned about in prompts crafted by teachers could be present in prompts crafted by crowds, raising questions about how the specific positions expressed in a prompt might interact with future contributors' own positions. In the context of our meta-talk task, we imagine that a worker asked to improve a discussion prompt that represented a position they are opposed to might change the sentiment, the question, or even the topic of the prompt. We also expect this to happen more generally in iterative writing tasks: people asked to engage with text that contradicts their own positions might find it hard to do so. For instance, in *Wikum*, a crowd-writing system designed to summarize comment threads, participants assigned to summarize a political discussion reported feeling that, "[...] summarizing content they disagreed with took more effort" [230, pg. 2090].

Our task design addresses these considerations in several ways. First, in some conditions, we provide explicit criteria around listening, comparing, and questioning as described above, to allow us to study the effects of providing clear guidance. In others, rather than asking workers too write the prompt as a whole, we ask them to write separate sentences to address each criterion, with the thought that focusing attention on each of the three sub-tasks would help. Second, we surveyed each participant before and after the task to consider how having a neutral versus a non-neutral initial position on the topic may affect their performance. Third, we vary the positions expressed in the comments people are crafting a prompt around to explore how positional agreement affected people's ability to successfully complete the task.

6.4 Study Design: Crafting Discussion Prompts

6.4.1 Research Questions

Based on the above discussion, we designed an experiment to address a number of questions around the use of meta-talk tasks as a tool for onboarding newcomers to an online policy discussion.

RQ1. To what extent can newcomers to a discussion effectively create meta-talk-based discussion prompts?

- **1a**. How do people perform when writing a new policy discussion prompt versus improving an existing one?
- **1b**. How does the structure of the task affect performance? How much do instructions and explicit subdivision of the task into sub-tasks improve people's ability to complete the task?
- **1c**. How does the position toward the policy proposal of the selected comments affect performance, in terms of both their relationship to each other and to a newcomer's own position toward the proposal?
- **1d**. Apart from task structure and policy position, when and why do people tend to perform well or poorly at the task?

RQ2. To what extent does crafting meta-talk discussion prompts affect newcomers' subsequent contribution to the discussion, in terms of their engagement with others, topicality, reasoning, and effort, compared to performing baseline onboarding tasks?

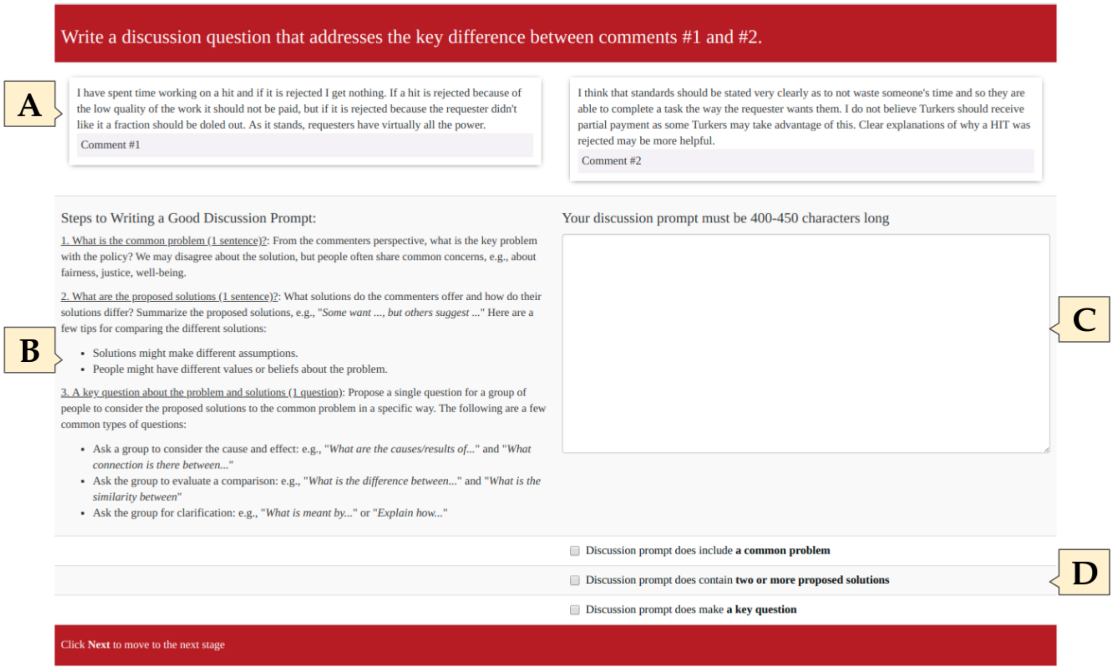
In the experiment, Amazon Mechanical Turk (AMT) workers were invited, as newcomers, to an online policy discussion about the AMT participation agreement. Previous studies have shown that requesters rejecting work is a thorny issue [103, 147] and that AMT workers have proposed to amend the policy to provide them with either partial payment for the work they had completed or a second chance to fix their work [152]. We designed a study that simulated newcomers entering an online policy discussion around the Partial Payment proposal.

6.4.2 Onboarding Activity Conditions

Variations of the main meta-talk prompt task

To address our questions about newcomers' ability to write prompts, we developed a number of variations on an onboarding task that involved crafting a meta-talk discussion prompt based on comments drawn from a policy discussion. Participants in the meta-talk conditions were assigned to perform a *Prompt Task* (Write or Improve) with a particular *Design* (No Structure, Instructions, or Scaffolded) and *Content* (Biased or Balanced)—a 2x3x2 factorial between-subject design. Figure 6.1 presents a screenshot of the interface in the Write-Instructions-Biased condition, with labels indicating each experimentally varied component. Below we discuss in more detail how they varied.

Prompt Task (Write, Improve): In the Write condition, participants were presented with an empty text box and were asked to "Write a discussion question that addresses the key difference between comments #1 and #2." In the Improve condition, participants were presented with one of three prompts that had been



Write a discussion question that addresses the key difference between comments #1 and #2.

I have spent time working on a hit and if it is rejected I get nothing. If a hit is rejected because of the low quality of the work it should not be paid, but if it is rejected because the requester didn't like it a fraction should be doled out. As it stands, requesters have virtually all the power.

Comment #1

I think that standards should be stated very clearly as to not waste someone's time and so they are able to complete a task the way the requester wants them. I do not believe Turkers should receive partial payment as some Turkers may take advantage of this. Clear explanations of why a HIT was rejected may be more helpful.

Comment #2

Steps to Writing a Good Discussion Prompt:

1. What is the common problem (1 sentence)?: From the commenters perspective, what is the key problem with the policy? We may disagree about the solution, but people often share common concerns, e.g., about fairness, justice, well-being.

2. What are the proposed solutions (1 sentence)?: What solutions do the commenters offer and how do their solutions differ? Summarize the proposed solutions, e.g., "*Some want ..., but others suggest ...*" Here are a few tips for comparing the different solutions:

- Solutions might make different assumptions.
- People might have different values or beliefs about the problem.

3. A key question about the problem and solutions (1 question): Propose a single question for a group of people to consider the proposed solutions to the common problem in a specific way. The following are a few common types of questions:

- Ask a group to consider the cause and effect: e.g., "*What are the causes/results of...*" and "*What connection is there between...*"
- Ask the group to evaluate a comparison: e.g., "*What is the difference between...*" and "*What is the similarity between*"
- Ask the group for clarification: e.g., "*What is meant by...*" or "*Explain how...*"

Your discussion prompt must be 400-450 characters long

□ Discussion prompt does include **a common problem**

□ Discussion prompt does contain **two or more proposed solutions**

□ Discussion prompt does make **a key question**

Click **Next** to move to the next stage

Figure 6.1: Screen shot of the Write-Instructions interface. **A**. Participants were provided two comments as a reference. In the Improve task condition, the comments were accompanied by an existing prompt to improve. **B**. Steps to writing a good discussion prompt, provided in the Instructions and Scaffolded task layout conditions. **C**. Text area with a 400-450 character restriction. In the Scaffolded condition, the interface included three additional text boxes for each of the steps to writing a good discussion prompt. **D**. Self-assessment of the discussion prompt via check box items, which varied by the presence or absence of Instructions.

crafted by participants in the Write-Scaffolded-Balanced condition, then asked to "Improve the discussion question to address the key difference between comments #1 and #2." In both conditions, we restricted the submission text box to prompts between 400-450 characters, to reduce the effect of prompt length on acceptability, by imposing a standard. As we did not test an *Evaluate* condition, four researchers reviewed prompts from the Write-Scaffolded-Balanced condition to select the following by consensus for the *Improve* condition:

- *Prompt 1*: Defining a task clearly so workers will comprehend the requirements is the common problem. One solution proposes clearer standards for work, while another suggests both objective standards and partial payment. How can diverse requesters and tasks be made to hold to an objective standard of task clarity? Defining different levels of task difficulty could be helpful to workers in deciding if they will work on any particular task.
- *Prompt 2*: The concern seems to be the lack of appeal for workers, stemming from lack of communication between workers and requesters. One commenters suggests partial pay for rejected HITs, but the other says that better communication as to HIT guidelines and rejection reasons would solve the issue. What can be done on the Mturk platform to better facilitate communication about the requeters HITs from the task description to their acceptance/rejection?
- *Prompt 3*: There are concerns about the power balance between turkers and requesters, and how rejected hits are dealt with. Some would like rejected hits to be accompanied with a clear explanation for the rejection, while others feel they should be compensated for what they correctly did. Should rejected hits be viewed as an opportunity to learn from and improve ones turking, or should requesters seek to improve their own hits?

Task Design (No Structure, Instructions, Scaffolded): We varied the Write and Improve prompt task design in three ways. Participants in the No Structure condition were provided with the basic instructions listed in the Prompt Task description (above). Participants in the Instructions condition were also presented with a set of *Steps to Writing a Good Discussion Prompt* based on existing guidelines [98, 177] (shown in Figure 6.1B):

1. *What is the common problem* (1 sentence)? From the commenters perspective, what is the key problem with the policy? We may disagree about the solution, but people often share common concerns, e.g., about fairness, justice, well-being.
2. *What are the proposed solutions* (1 sentence)? What solutions do the commenters offer and how do their solutions differ? Summarize the proposed solutions, e.g., "Some want ..., but others suggest ..." Here are a few tips for comparing the different solutions: Solutions might make different assumptions, people might have different values or beliefs about the problem.
3. *A key question about the problem and solutions* (1 question) Propose a single question for a group of people to consider the proposed solutions to the common problem in a specific way. The following are a few common types of questions: e.g., "What are the causes/results of..." "What connection is there between..." and "What is meant by..."

Participants in the Scaffolded condition saw these instructions, but instead of writing the discussion prompt as a whole, they assembled it by writing 1-2 sentences for each step, using separate textboxes for each step. The Scaffolded task was inspired by similar crowd-writing tasks [110, 119, 230]. In the Instructions and Scaffolded conditions participants were presented with three checkboxes corresponding to each of the steps to writing a good discussion prompt (Figure 6.1D), inspired by the observation by Dow et al. [53] that tasking participants to self-assess their work improved performance. In the No Structure condition participants were presented with a checkbox with the message: "Discussion prompt does not need further improvement."

Task Content (Biased, Balanced): The two specific reference comments on which participants were asked to base their discussion prompt could have a significant effect on performance. We therefore varied these two comments, drawing them from a set of three where two supported and one opposed Partial Payment:

- *Support-PP1*: "I feel that the requester should communicate more with the turker. I agree with the partial payment, there should be some kind of compensation given for the time spent on certain types of hits. With a transcription, there could be some small errors that the requester was not happy about, but the overall time spent was too great to ignore."
- *Support-PP2*: "I have spent time working on a hit and if it is rejected I get nothing. If a hit is rejected because of the low quality of the work it should not be paid, but if it is rejected because the requester didn't like it a fraction should be doled out. As it stands, requesters have virtually all the power."
- *Oppose-PP*: "I think that standards should be stated very clearly as to not waste someone's time and so they are able to complete a task the way the requester wants them. I do not believe Turkers should receive partial payment as some Turkers may take advantage of this. Clear explanations of why a HIT was rejected may be more helpful."

The Balanced condition included exposure to a support-oppose pair; the Biased condition included exposure to a support-support pair. Twenty participants were assigned to each Content pair variation in the Meta-talk conditions.

Baseline onboarding task conditions.

Two other activities served as baselines to compare the effects of doing the meta-talk activity on subsequent contributions to the discussion forum.

AMT Policy Baseline Task. Participants assigned to this baseline condition were presented with an excerpt from the AMT Participation Agreement and were asked to propose three sentences to delete. After selecting three sentences, the participant was asked to respond to the following prompt: "Why are these the right sentences to remove from the AMT Participation Agreement?"

Active Listening Baseline Task. Participants assigned to this baseline condition were presented with a pair of existing comments drawn from the set of three used for the meta-talk task. Using an interface modeled after the *Reflect* platform [119], participants were asked to highlight a phrase in each comment that captures the commenter's position on the policy and then asked to write what they "[...] hear this commenter saying."

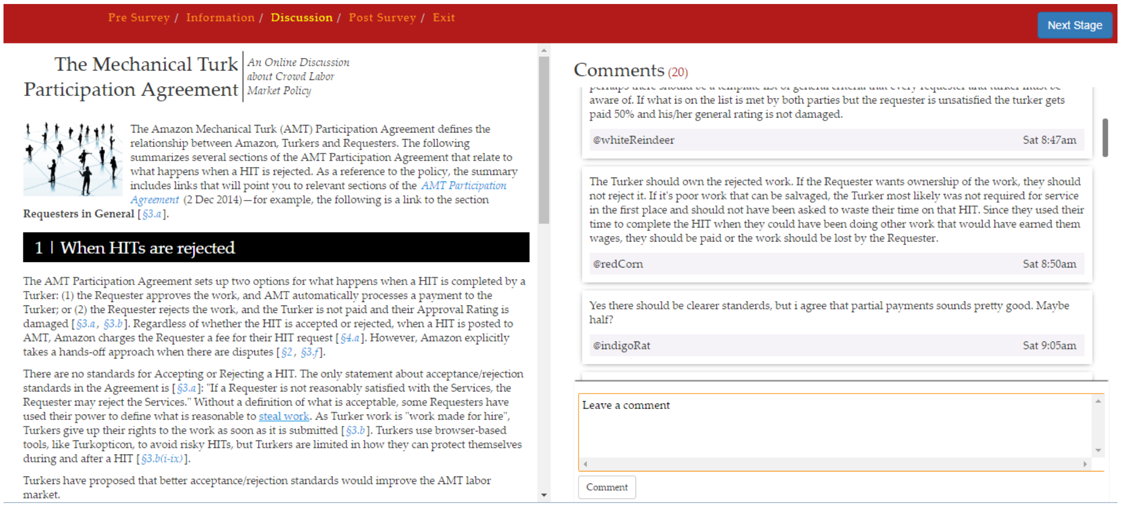
6.4.3 Recruitment and Procedure

The HIT description recruited Turkers to test the user interface of a novel online discussion forum system. We did not restrict access to the HIT (e.g., to Turkers from specific countries or with specific levels of experience); however, a majority of participants were U.S.-based Turkers who spoke English as a first language. All participants who completed the HIT were rewarded $4.50 for their time; the average time to completion was 30 minutes, which equates to an hourly rate of $9.

We received IRB-approved informed consent from all participants. While the HIT was active, we monitored Turkopticon [103] and TurkerNation [147] closely to listen for any problems related to the HIT, in addition to any concerns we received directly through the AMT interface. We also maintained an active "dummy HIT" to compensate Turkers who encountered a technical error with the system. Finally, we indicated that our research group is not associated with Amazon and that the purpose of the experiment was for research. We applied these actions based on recommended best practices for ethical conduct of academic research with crowd workers [151].

Prior to entering the experiment, participants were informed that they would be participating in a discussion about the policy topic "what should happen when a HIT is rejected?" and that they would have an opportunity to add their voice to the discussion. We informed participants that the intent of the discussion is to help resolve a lack of consensus among Turkers around two proposals to address the question: partial payment and second chance. Before starting the experiment, participants were asked to rate their initial preference toward both the partial payment and second chance proposals on a 5-item scale, from strongly disagree to strongly agree. They were also asked basic demographic questions and questions about their Turking experience, variables we used as control characteristics in our statistical analyses.

Participants were then randomly placed into one of the fourteen conditions to complete before entering the discussion forum. The discussion interface was modeled after RegulationRoom, a platform for civic engagement in public rulemaking [185]. The interface included two side-by-side panels (Figure 6.2). The left panel included a summary of the AMT Participation Agreement policy and



The Mechanical Turk Participation Agreement | *An Online Discussion about Crowd Labor Market Policy*

The Amazon Mechanical Turk (AMT) Participation Agreement defines the relationship between Amazon, Turkers and Requesters. The following summarizes several sections of the AMT Participation Agreement that relate to what happens when a HIT is rejected. As a reference to the policy, the summary includes links that will point you to relevant sections of the *AMT Participation Agreement* (2 Dec 2014)—for example, the following is a link to the section **Requesters in General** [*§3.a*].

1 | When HITs are rejected

The AMT Participation Agreement sets up two options for what happens when a HIT is completed by a Turker: (1) the Requester approves the work, and AMT automatically processes a payment to the Turker; or (2) the Requester rejects the work, and the Turker is not paid and their Approval Rating is damaged [*§3.a*, *§3.b*]. Regardless of whether the HIT is accepted or rejected, when a HIT is posted to AMT, Amazon charges the Requester a fee for their HIT request [*§4.a*]. However, Amazon explicitly takes a hands-off approach when there are disputes [*§2*, *§3.f*].

There are no standards for Accepting or Rejecting a HIT. The only statement about acceptance/rejection standards in the Agreement is [*§3.a*]: "If a Requester is not reasonably satisfied with the Services, the Requester may reject the Services." Without a definition of what is acceptable, some Requesters have used their power to define what is reasonable to steal work. As Turker work is "work made for hire", Turkers give up their rights to the work as soon as it is submitted [*§3.b*]. Turkers use browser-based tools, like Turkopticon, to avoid risky HITs, but Turkers are limited in how they can protect themselves during and after a HIT [*§3.b(i-ix)*].

Turkers have proposed that better acceptance/rejection standards would improve the AMT labor market.

Comments (20)

perhaps there should be a template list of general criteria that every requester and turker must be aware of. If what is on the list is met by both parties but the requester is unsatisfied the turker gets paid 50% and his/her general rating is not damaged.

@whiteReindeer Sat 8:47am

The Turker should own the rejected work. If the Requester wants ownership of the work, they should not reject it. If it's poor work that can be salvaged, the Turker most likely was not required for service in the first place and should not have been asked to waste their time on that HIT. Since they used their time to complete the HIT when they could have been doing other work that would have earned them wages, they should be paid or the work should be lost by the Requester.

@redCorn Sat 8:50am

Yes there should be clearer standerds, but i agree that partial payments sounds pretty good. Maybe half?

@indigoRat Sat 9:05am

Figure 6.2: Screen shot of the online policy discussion interface.

proposals to amend it. The language for the summary, as well as the partial payment and second chance proposals associated with the deliberation "What should happen when a HIT is rejected?", were developed through a prior study [154].

The right panel included a discussion forum interface and a text box where participants could enter their comments. To populate the discussion thread, we selected 20 comments contributed in the prior study, half in favor of and half opposed to partial payment. We added time stamps to make the discussion appear recent and assigned participants a pseudonym (a concatenated color and animal, e.g., *@purpleOstrich*) to each comment in the thread. The interface did not include voting or other mechanisms for engaging with the content (e.g., reply, like, or share) as our focus was specifically on commenting behavior.

Participants received the following instructions when transitioning to the discussion forum: "You are about to begin the Discussion portion of the task. After 1 minute you will be able to leave the discussion and move onto the post-survey; however, you are welcome to remain in the online discussion as long as the HIT permits. Payment for this HIT is entirely based on your pre- and post-survey responses.” Although participants were only required to spend one minute testing the discussion forum interface, the average dwell time was 4 minutes, 35 seconds (SD 3:37).

6.5 Data Analysis

6.5.1 Coding for Prompt Task and Comment Quality

We evaluated the discussion prompts based on criteria from the *Steps to writing a good discussion prompt* presented in section 6.4.2. Evaluations were conducted by two researchers, each training on a sample of 120 prompts and testing on a held out set of 65. Table 6.1 presents the Cohen’s kappa scores for inter-rater reliability at each level of the evaluation criteria.

- **Listen**: (binary) The discussion prompt correctly identifies the common problem.
- **Compare**: (binary) The discussion prompt presents a comparison of the different views.
- **Question Type**: (“None”, “Rhetorical”, “Genuine”) The type of question asked, if any.
- **Acceptable**: (binary) Acceptable discussion prompts (“True”) are those that include Listening, Comparing, and a Genuine question.

The discussion comments were coded for their expressed position (e.g., agreement or disagreement with partial payment and second chance) and whether they presented some form of reasoning for their position. Because a comment might include a mix of positions, we coded for both partial payment and second chance separately. We also coded for three types of commenting behaviors that signify engagement with a policy discussion [207]. To identify whether a comment was Topic Coherent, we coded for whether the topic of the comment added to the topics present in the first three comments closest to the text box [154, 208]. We coded whether a comment included a genuine or rhetorical question, or no question [207]. Finally, although the discussion forum interface did not include a "reply" feature, we coded for whether participants replied to another commenter by mentioning a user's pseudonym (e.g., *@purpleOstrich*), statements like "I agree with you," and evidence of linguistic entrainment (adopting the language of others).

6.5.2 Control Variables

To account for characteristics about the participant that might relate to performance of the prompt task or to participation in the discussion forum, we incorporated a series of control variables surveyed before a participant was exposed to the experimental conditions.

As the bias of a moderator is an important concern when crafting a policy discussion prompt [177, 202], we pre-surveyed participants on their initial preferences toward the partial payment and second chance proposals. To evaluate the influence of having a position versus not having a position, we coded a

	Train	*Test*	*Observations*	
Prompt Task Evaluation				
Listen	0.82	0.92	215	(66.9%)
Compare	0.92	0.90	211	(65.7%)
Question Type	0.81	0.85	229	(71.3%)
Comment position				
Partial Payment	0.86	0.84	190	(39.0%)
Second Chance	0.92	0.91	146	(29.9%)
Reasoned Opinion	0.85	0.72	236	(48.7%)
Discussion Comments				
Topic Coherent	0.81	0.78	187	(38.3%)
Question Type	1.00	0.85	23	(4.7%)
Reply	0.75	0.71	24	(4.9%)

Table 6.1: Report of the Cohen's Kappa score for inter-rater reliability, along with number and percentage of comments in which each code was observed. Two researchers trained with the coding scheme on 120 prompts and 120 comments until an acceptable level of inter-rater reliability was reached (Cohen's Kappa ≥ 0.7) and then tested on a held out set of 65 prompts and 65 comments. A total of 321 discussion prompts were crafted during the meta-talk onboarding task and 484 comments were posted to the discussion forum.

binary measure of whether participants rated their preference as "Neutral" or took a stance (either "Agree" or "Disagree") about partial payment.

In group situations, social sensitivity—a measure of how well a person works with others—is associated with positive performance on group discussion-based tasks. While we did not survey for social sensitivity, women tend to score better on the measure than men, and groups with a higher proportion of female participants tend to exhibit a higher collective intelligence than other groups when discussion is involved in the problem-solving [224]. For this reason, we include the participant's stated gender identity (Not disclosed, Female, Male). Because workers have different levels of investment in AMT and this might affect their performance or their position on partial payment, we also surveyed participants about their time spent and daily income from AMT.

At the discussion comment level, longer comments might be more likely to include characteristics that are valuable in policy deliberation such as reasoned opinions and topic coherence. For that reason, we also control for the character length of comments.

6.5.3 Statistical Models

To evaluate how prompt task, design, and content affect prompt acceptability, we used a standard logistic regression to measure the binary variable of prompt acceptability, as well as each individual element of the acceptability criteria (i.e., Listen, Compare, Question, Genuine question). Model significance was evaluated using the log-likelihood ratio test to compare the goodness-of-fit of a model incorporating just the control characteristics with a model that also includes experimental variables. This procedure provides both a higher threshold than comparison with just the intercept and consistency in treatment of the control characteristics through the modeling process.

The model coefficients are interpreted as the expected change that each independent variable contributes to the logits of the response variable. Throughout the findings, we exponentiate the logits to present the odds ratios, which can be interpreted as the change in the response variable expected from a one-unit increase of an independent variable, holding all others constant. For example, in the Write Prompt condition Question model (Table 6.4), the independent variable "Gen: Female" indicates that female participants are 3.28 times ($p \leq 0.001$) more likely than male participants to write prompts that meet the Question criteria.

As a few participants contributed more than one comment, we treated Participant as a mixed-effects nesting variable to account for non-independence when examining the effect of the onboarding activity on later participation in the discussion forum. Mixed-effects logistic regressions were used to predict the presence of both a reasoned opinion and topic coherence at the comment level. As with the standard logistic regression to model prompt acceptability, we exponentiated the logit estimate for each coefficient to present the odds ratios for each response variable.

As each model incorporates several independent and control variables, to account for the effect of multiple comparisons, we applied a Tukey-based post-hoc comparison of the estimated marginal means (EMMs) at each level of a significant factor variable. Through this procedure, the effect and significance of a variable are averaged over the effect and significance of other variables in a model to provide a more accurate assessment of its impact.

6.6 Findings

6.6.1 Descriptive Overview

Nearly half of those who started the task (988 participants) completed it (453). Those who completed were on average 34 years of age (SD 10 years) and about half identified as female (226 male, 219 female, 8 not disclosed). A majority of participants were located in the United States (85%) and speak English as a first language (92%). They reported earning about $15 per day (SD $13), doing tasks a few days per week for about 5 hours (SD 1.5) on average.

Baseline	Participants		
AMT Policy	56		
Active Listening	59		
Total	115		
Meta-talk	*Prompt Task*		
Design	Write	Improve	*Total*
No Structure	58	57	115
Instructions	50	56	106
Scaffolded	59	58	117
Total	167	171	338

Table 6.2: Breakdown of participants assigned to each of the onboarding conditions.

Table 6.2 shows the breakdown of participants by condition. Almost everyone who completed the task posted one or more comments in the following discussion (431 commenters, 484 total comments). Participants spent 4 minutes, 35 seconds in the discussion on average (SD 217 seconds), posting comments that averaged 297 characters (SD 226). Of the 338 participants in the meta-talk conditions, 321 attempted to write a prompt. To report quotes from the meta-talk task prompts, we assigned participants a unique ID ranging from P1-P321.

6.6.2 RQ1. Effects on Acceptability of Discussion Prompts

RQ1a. Prompt task

Table 6.3 reports on the overall and specific acceptability criteria for prompts in the Write and Improve conditions. In the Write conditions, approximately 38% of the Write prompts were rated overall acceptable ("True"), i.e., showed evidence of listening to points of connection, comparing the two comments, and asking a genuine question about them. In the Improve conditions, 50% of the prompts were rated acceptable, meaning that half of the initially acceptable prompts fed through an Improve task *declined* in quality.

RQ1b. Task design

Tables 6.4 and 6.5 report models of how the task design and content relate to performance on the Write and Improve tasks, respectively. When tasked to Write a prompt, the design of the task affected acceptability. Prompts written in the Instructions and Scaffolded task designs were 6.7 and 4.23 times, respectively, more likely to be Acceptable than those written in the No Structure design. The story is similar for the Listen and Question evaluation criteria: providing instructions and scaffolding for the task contributed to people's ability to meet the criteria. We did not find an effect of task design on prompt acceptability in the Improve prompt task.

	Acceptable		(A) *Listen*		(B) *Compare*		(C) *Question*		
Prompt	True	False	True	False	True	False	None	Rhetorical	Genuine
Task									
Write	62	101	99	64	104	59	55	21	87
Improve	80	78	116	42	107	51	39	24	95

Table 6.3: Prompt acceptability by task (i.e., Write, Improve), both overall and on specific evaluation criteria.

Write Prompt	*Acceptable* OR (SD)	*Listen* OR (SD)	*Compare* OR (SD)	*Question* OR (SD)	*Genuine* OR (SD)
(Intercept)	0.10 (0.5) ***	0.62 (0.4)	0.66 (0.4)	0.73 (0.4)	3.97 (0.7) *
D: Instructions	6.70 (0.5) ***	3.51 (0.4) **	1.77 (0.4)	3.98 (0.5) **	1.28 (0.7)
D: Scaffolded	4.23 (0.5) **	5.59 (0.4) ***	1.40 (0.4)	4.53 (0.4) ***	1.06 (0.6)
C: Balanced	1.58 (0.4)	0.62 (0.4)	2.38 (0.4) *	0.45 (0.4) .	1.50 (0.5)
Control Characteristics					
PP: Neutral	2.09 (0.4) .	1.25 (0.4)	1.81 (0.4)	2.15 (0.5)	1.64 (0.6)
Daily Earn	0.89 (0.2)	0.97 (0.2)	1.03 (0.2)	0.86 (0.2)	1.03 (0.3)
Gen: Female	1.35 (0.4)	1.67 (0.4)	0.93 (0.4)	3.28 (0.4) **	0.47 (0.5)
Gen: ND	1.64 (1.1)	2.13 (1.2)	1.99 (1.2)	1.80 (1.2)	—
Goodness-of-fit $\chi^2(3, 163)$=22.2 ***		=20.4 ***	=8.3 *	=18.0 ***	=0.8

Table 6.4: Acceptable prompts in the Write task condition, by Design (i.e., No Structure, Instructions, Scaffolded) and Content (i.e., Biased, Balanced). To evaluate the goodness-of-fit of each model to the data, we used the log-likelihood ratio test to compare a model with just the Control Characteristics against the full model and report the χ^2 statistic. Note that including the condition variables did not significantly improve the fit of the *Genuine* model to the data. p-value significance codes: 0.0001 '***' 0.001 '**' 0.01 '*' 0.05 '.'

Improve Prompt	*Acceptable* OR (SD)		*Listen* OR (SD)		*Compare* OR (SD)		*Question* OR (SD)		*Genuine* OR (SD)	
(Intercept)	0.37 (0.4)	*	1.22 (0.4)		1.59 (0.4)		1.42 (0.4)		1.15 (0.5)	
D: Instructions	1.66 (0.4)		1.75 (0.4)		0.83 (0.5)		1.32 (0.5)		3.31 (0.7)	
D: Scaffolded	0.56 (0.4)		0.89 (0.4)		0.47 (0.4)	.	0.80 (0.5)		0.89 (0.6)	
P: Prompt 2	3.98 (0.4)	**	2.53 (0.4)	*	1.94 (0.4)		2.38 (0.5)	.	5.61 (0.6)	**
P: Prompt 3	5.09 (0.4)	***	3.28 (0.4)	**	3.20 (0.4)	**	2.69 (0.5)	*	10.20 (0.6)	***
Control Characteristics										
PP: Neutral	0.95 (0.4)		0.73 (0.4)		1.03 (0.4)		0.77 (0.4)		1.17 (0.6)	
Daily Earn	0.82 (0.2)		0.92 (0.2)		1.11 (0.2)		0.95 (0.2)		0.64 (0.2)	*
Gen: Female	1.12 (0.3)		1.34 (0.4)		1.03 (0.4)		1.84 (0.4)		0.62 (0.5)	
Gen: ND	—		—		—		—		—	
Goodness-of-fit χ^2(4, 158)	=21.51	***	=9.49	*	=9.81	*	=6.39		=18.04	**

Table 6.5: Acceptable Prompts in the Improve task condition, by Design (i.e., No Structure, Instructions, Scaffolded) and Prompt (i.e., which specific prompt was being improved). As in Table 6.4, we report the χ^2 statistic from a goodness-of-fit log-likelihood ratio test of each model. Note that participants in the Improve-Scaffolded task were less likely to meet the Compare criteria, though this was not significant in a Tukey-based post-hoc comparison. p-value significance codes: 0.0001 '***' 0.001 '**' 0.01 '*' 0.05 '.'

RQ1c. Content and position toward the policy proposal

The models in tables 6.4 and 6.5 also show the relationship between content and prompt acceptability in the Write and Improve tasks, respectively. When tasked to Write a prompt, the effect of having an initial position toward the policy on prompt acceptability was marginally significant, such that people who came in with pre-existing positions were less likely to craft an acceptable prompt (Table 6.4, "PP: Neutral"). Whether the reference comments were balanced or biased did not affect overall prompt acceptability, but it did affect participants' ability to write a prompt that compares them (Table 6.4, "C: Balanced"). When the comment positions were balanced with one in support and the other in opposition to Partial Payment, participants were 2.38 more likely to write a prompt that includes a comparison between the comments than if the comments are biased toward support.

In the Improve task, the content of the prompt to improve affected acceptability (Table 6.5). Specifically, improvements to Prompt 2 and Prompt 3 were 3.98 and 5.09 times, respectively, more likely to be acceptable than improvements to Prompt 1, suggesting there was something about the content of Prompt 1 that induced people to deviate from the instructions; we discuss this below.

RQ1d. When and why people tend to perform well or poorly

In both the Write and Improve tasks, some control characteristics are significant predictors of acceptability. Female participants are more likely to meet the Question criteria in the Write task, though this effect is not significant in the Improve task. Participants who report earning less per day on AMT were more likely to transform the provided prompt in the Improve task from a genuine to a rhetorical question than participants who earn more.

We next look more closely at the specific errors people made. In the Write task, participants in the No Instructions condition rarely wrote acceptable prompts (only 16%), so we focus on the unacceptable prompts written by people in the Instructions and Scaffolded conditions with the goal of discovering kinds of errors people made even when given useful instruction. To do this, we manually reviewed the text of the unacceptable prompts in those conditions. This was not a formal content analysis coding process; we simply read through the unacceptable prompts and looked for representative examples of the errors that would be useful for other researchers and system designers to think about.

One pattern was the presence of personal opinion. Some participants ignored the separate components to writing a prompt and just offered an opinion. Examples include (P142) "[...] I think amazon should have some rules in places that help protected it's turkers [...]" and (P308) "[...] small tasks with set answers need to be checked before issuing a rejection. We could propose an automatic system when the tasks are big enough to catch mistakes." Many other participants provide some acceptable prompt components, but also a personal opinion, as in the next example that identifies a common problem and explicitly compares the comment pair, yet offers a personal opinion where there should be a question in the last sentence (P261):

> "Both the comments adressing HIT rejections. Comment #1 says reason to get paid for rejected work, based on the quality of the work. Where comment #2 says Requesters must communicate with turkers upon rejection, and make a partial payment for the time spent on the

> work. Whether there is payment made or not made, but rejections without fair reasons or just rejections made without looking quality of work, would really hurt a turker."

Other prompts imply an opinion by the way their question was phrased. The following are rhetorical questions posed by otherwise acceptable prompts:

- (P19) "How do we know why or how our work was rejected if the final say is all up to the requester?"
- (P87) "How often are undue rejections received for lengthy time consuming work any way?"
- (P116) "But doesn't doling out partial payment create more hassel for the requestor?"
- (P319) "Does the requestor not see it as fair for the worker to recieve either a [SC or PP]?"

Errors in the Improve task were largely due to the question in the provided prompt. By reviewing the less acceptable improvements, we realized that the question in Prompt 1 could be seen as being rhetorical or narrow: *"How can diverse requesters and tasks* ***be made to hold to*** *an objective standard of task clarity?"* The bold words imply holding Requesters accountable, rather than other means of improving task clarity. This meant that participants exposed to Prompt 1 had the additional challenge of converting the question to a less pointed form, such as (P111) "How can the requesters improve their standards of clarity?" or to a more open and genuine question (P309) "What are some ways in which a requester can help workers complete their HITs and get better quality work?"

The question also makes an assumption that task clarity is the *right* solution, which invited opposition, such as (P276) “Will stronger guidelines really help?” and (P209) “If the requester states their instructions were clear but still rejected the work, who is at fault then?”

6.6.3 RQ2. Effects of a Meta-talk task on Discussion Comments

We now turn to our second main research question, about whether performing the prompt construction task affects participants’ subsequent commenting behavior. As mentioned earlier, most (431 of 453) participants posted a comment, and we found no significant difference in the likelihood to comment by condition, χ^2(7, N=459) = 8.172, p = 0.31.

We had planned to look at a range of response variables, including whether participants asked genuine questions, engaged directly with other posters, provided reasoned opinions, and posted comments that were topically coherent with the existing discussion. However, fewer than 5% of the comments included questions or replies (Table 6.1), so we could not build meaningful statistical models predicting those variables. The response variables we were left with—reasoned opinion and topic coherence—were not significantly predicted by task or design (Table 6.6). Specifically, longer comments are more likely to include a reasoned opinion and our attempts to model factors that predict topically-coherent comments were largely unsuccessful.

Comment length can be taken as an indicator of effort, and longer comments, because they have more exposition, are also more likely to contain elements such as reasoned opinion. Thus, we next looked at what variables might predict

	Reasoned Opinion		*Topic Coherence*	
Discussion Comments	OR (SD)		OR (SD)	
(Intercept)	0.80 (0.27)		0.47 (0.29)	**
T: Improve	1.04 (0.24)		1.08 (0.24)	
D: Instructions	0.61 (0.31)		1.66 (0.31)	
D: Scaffolded	0.93 (0.30)		0.88 (0.31)	
Control Characteristics				
Acceptable: Listen	1.46 (0.27)		1.06 (0.27)	
Comment Length	2.09 (0.15)	***	1.08 (0.12)	
Goodness-of-fit: $\chi^2(5, 367)$ =41.39 ***			=5.78	

Table 6.6: Reasoned Opinion and Topic Coherence by exposure to the Prompt task (i.e., Write, Improve), Design (i.e., No Structure, Instructions, Scaffolded) and by the prompt task Listen evaluation criterion.
p-value significance codes: 0.0001 '***' 0.001 '**' 0.01 '*' 0.05 '.'

	Write		*Improve*	
Discussion Comments	Exp. Char. (SD)		Exp. Char. (SD)	
(Intercept)	292.83 (0.11)	***	309.87 (0.10)	***
B: Active Listening	0.92 (0.14)		0.93 (0.13)	
D: No Structure	1.35 (0.14)	*	1.29 (0.13)	*
D: Instructions	1.06 (0.15)		0.91 (0.13)	
D: Scaffolded	0.87 (0.14)		1.00 (0.13)	
Control Characteristics				
PP: Neutral	1.02 (0.10)		1.01 (0.09)	
Daily Earn	1.04 (0.05)		1.06 (0.04)	
Gen: Female	1.16 (0.09)	.	1.03 (0.08)	
Gen: ND	1.43 (0.30)		1.08 (0.33)	
Goodness-of-fit: $\chi^2(4, 282)$ =12.73 **			=10.57 *	

Table 6.7: Total expected discussion comment length ("Exp. Char.") by exposure to the Baseline (i.e., AMT Policy, Active Listening) and Meta-talk design conditions.
p-value significance codes: 0.0001 '***' 0.001 '**' 0.01 '*' 0.05 '.'

comment length. To do this, we applied a negative binomial regression, which is appropriate for modeling count variables that are over-dispersed, meaning that the variance is greater than the mean. We exponentiated the coefficients from the negative binomial regression to present the incident rate ratios associated with each independent variable and expected comment length.

In the best-fitting models we could construct, the Write-No Structure and the Improve-No Structure tasks led to 1.35 and 1.29 times longer discussion comments, respectively, than the Baseline AMT Policy task (Table 6.7, "D: No Structure" line). A Tukey-based pairwise comparison between No Structure and the other design conditions showed that participants were more likely to write longer comments when exposed to the Write-No Structure versus the Write-Scaffolded design (1.53 times, p≤0.01), or when exposed to the Improve-No Structure condition versus the Improve-Instructions condition (1.42 times, p≤0.03).

Our interpretation of the findings about comment length is that effort on the task may have impacted effort in the discussion forum. The more involved meta-talk conditions, with instructions to consider and multiple steps to complete, may have consumed more of participants' time and effort, reducing the effort they put into commenting later.

6.7 Discussion and Limitations

In the related work, we define meta-talk as "talk about the talk" that aims to revisit points of conflict in an existing discussion. We argue that newcomers to a discussion could create opportunities for meta-talk by crafting policy discussion prompts that encourage existing members to talk about their different perspectives. Further, we leveraged ideas from crowd-writing (see section 6.3.4) to test parts of a task designed to help newcomers to build policy discussion prompts from two existing comments or to improve a prompt drafted by another newcomer.

The experiment to test this addressed two primary research questions (see section 6.4.1). Our first question was *to what extent can newcomers to a discussion effectively create meta-talk-based discussion prompts?* We split this question into four parts to examine how task structure (RQ1a), design (RQ1b), and policy position (RQ1c) relate to task performance, as well as when and why people tended to perform poorly (RQ1d). Inspired by the potential for a micro-task to affect discussion norms [119], we also asked *to what extent does crafting meta-talk discussion prompts affect newcomers' subsequent contribution to the discussion?* (RQ2)

To briefly summarize our findings, (RQ1a) writing and improving policy discussion prompts is a difficult task, but (RQ1b) one that newcomers can perform reasonably well with appropriate structuring (and some caveats described below). While we found little effect of having an initial policy position (RQ1c), we did find that complementary perspectives in the comment pair improved their ability to complete the Write task and aspects of the question in the draft prompt affected performance on the Improve task. Our analysis of common errors (RQ1d) revealed that the participants, being informed and opinionated, engaged with the task material by inserting their own experience and by challenging assumptions implicit in the comment pair or the draft prompt provided for the task. However, our hope that a newcomer onboarding meta-talk task would lead to more engaged comments in the discussion (RQ2) was not supported. Here we unpack how these findings contribute to the design of meta-talk-based tasks and to future work on onboarding newcomers to discussion communities.

6.7.1 Provide clear instructions and scaffolding

We were encouraged by people's performance on the *Write* task; almost 40% of the composed prompts met the goals for a meta-talk-based prompt including hearing, comparing, and writing a genuine question to encourage other discussants to engage with each other. This is not an easy task; the criteria are rigorous, and the costs of failure might be detrimental to a group policy discussion, much as a biased [202] or untrained facilitator [145, 61] can negatively impact the group. We found that clear guidance helped many participants accomplish the task despite its difficulty; those who received support for the task in terms of instructions and scaffolding were much more likely to write acceptable prompts (50%) than those who did not (16%).

Providing clear instructions aligns with best practices drawn from related crowd-writing systems [91, 119, 110] and effective crowd task design more broadly [114, 151]. However, designing good instructions is not easy, and in our case it took several iterations before arriving at the designs presented here. Looking back, the designs share many characteristics with rubrics and peer feedback systems designed for MOOCs [121]: the instructions align closely with the sub-criteria; they give definitions, considerations, and examples to illustrate each of the criteria; and the interface provides people with a self-assessment tool toward the criteria.

More generally, our results emphasize the magnitude of the need for effective instructions, as participants were three times better at crafting acceptable prompts given appropriate guidance. Taking the time to iteratively pilot task instructions is important, in terms of both quality of outputs and worker fairness, by reducing time-wasting and pay-reducing rejections induced by ambiguous instructions [103, 152].

6.7.2 Select appropriate content for comparison tasks

Our results also demonstrate that for complex synthesis and iteration tasks, designers should pay careful attention to the balance of perspectives exhibited by the content presented. For instance, in the *Write* condition, participants found it easier to effectively compare two comments whose positions were farther apart than two comments that were roughly in agreement. In this experiment, we hand-chose each pair of comments to support experimental control, but in a real system, topic modeling and sentiment analysis could be used to select comment sets that are more likely to have topical agreement, but differing opinions. Social network analysis techniques might also be used to mitigate the concern raised in the background section around newcomers inadvertently encouraging interaction between people with a negative relationship [14, 156].

Several crowd-writing systems also include a process for selecting sets of comments that exhibit different properties. For example, *Knowledge Accelerator* [91] implements a process for filtering a list of information down to a unique set with a comparison task where workers “are asked to sample random items from the data in order to create a set of non-matching items [...]". The task continues until "a worker's familiarity with the distribution gives them a sense that [the items in the list] represent substantively different topics" [91, pg. 2264]. We imagine this pattern, called "Open-ended Set Sampling," could be adapted to filter a thread of comments for exemplars of each policy perspective to identify the points of conflict. For other discussion goals, such as constructing meta-talk about opportunities for consensus or summarizing the arguments toward particular policy proposals, the open-ended set sampling could reduce a list of comments to those that exhibit a similar position (e.g., support for partial payment), but with different reasoning.

6.7.3 Consider evaluation as a way to facilitate discussion

Our finding that many prompts declined in quality in the *Improve* task raises questions about deciding both what prompts are worth iterating on and when to stop iterating. Many crowd-writing systems set a fixed number of iterations [9, 91, 134] or apply an agreement-based scheme [38, 54, 221] as the stopping criteria for an iterative process.

In the case of "acceptable" policy discussion prompts, deciding when to stop is harder even though the high level desirable properties of listening, comparing, and questioning are well-defined. As we saw, one of the prompts in the Improve condition had a question that, although considered acceptable in our initial coding, in retrospect felt more rhetorical and implied a more concrete policy solution than the others. Other work has shown that people opposed to a point of view often engage with it indirectly [154, 166], and we saw that participants exposed to that prompt tended to ask rhetorical questions that challenged its assumptions. Possible ways to reduce that tendency would be to ask workers to provide their reasoning to justify a proposed improvement [54] or to introduce an active listening step to facilitate writer-and-improver dialogue [119], which may have an added benefit of introducing newcomers to each other [116].

We did not evaluate the effectiveness of the policy discussion prompts in fomenting actual meta-talk. To do so would involve organizing a group to consider a prompt and then studying their discussion [13]. However, simply inserting a prompt into an ongoing discussion might not trigger the reflection or innovative thinking intended by newcomer-crafted meta-talk [30, 132, 207]. Instead, we might draw inspiration from the ways that crowd-writing systems manage progress toward a complex objective. For example, the *Mechanical Novel* implements a process to crowd-write a fictitious story, yet allowing the plot to deviate by periodically *reflecting* on current progress and then *revising* the intermediary goals. This process of "[...] looping between reflecting on progress to identify a goal and revising based on that goal, allows [a group of workers] to converse with their work and evaluate options by trying them out" [111, pg. 235]. In a similar way, the process of evaluating several newcomer prompts could be formalized as a periodic group activity for existing members, which would provide them with a way to reflect on and revisit older topics by considering prompts generated from new perspectives.

6.7.4 Limitations and Open Questions

The most salient limitation of this study is that we traded off the ecological validity of real discussions for experimental control. This leads to questions about how to apply and extend these findings when generalizing from Turkers to discussion newcomers, from this task design to actual forum designs, and from this one-time HIT to ongoing policy discussions.

We believe that choosing the AMT participation agreement as the policy context mitigates many of the concerns about how Turkers might be different from newcomers to policy discussions more generally. For instance, it is possible that having some background knowledge of the topic would improve people's ability to construct effective prompts, and newcomers to a discussion will often have a pre-existing interest in the topic. Choosing Turkers and the AMT participation agreement is representative of this situation—so much so that the question of whether Turkers could construct prompts for an arbitrary discussion topic as part of a generic crowd-powered workflow is still open. We think it also reduces concerns that Turkers' paid status affects their motivation toward the task relative to volunteer newcomers, since Turkers do have a stake in the participation agreement. Still, pay matters, and whether newcomers to a policy discussion would be willing to spend the time required to do the onboarding tasks is also an open question, although cases like Reflect [119] suggest the answer is sometimes yes.

Another limitation is that the experimental design did not integrate the onboarding task and the discussion forum, which were posed as different stages/sub-tasks in the larger context of a HIT and had different visual designs and no direct incorporation of materials from one task to the other. We suspect this reduced possible carryover of socialization effects from doing the onboarding task into the discussion. More generally, in this experiment we did not consider how to integrate the prompts later into the discussion or how participants would interact with the prompts because our focus was on the support of prompt creation and newcomers' initial forum behavior. The thoughts around prompt evaluation above are a start, but the question of how to use newcomers' inputs in a facilitator's, moderator's, or group's larger policy discussion process is still unresolved.

Finally, our choice of a controlled, simulated, one-shot discussion leaves open questions about more natural and continued engagement. Because much participation in online deliberations comes from one-time contributors [153], we think many of the findings are likely to hold up for many participants in real discussions. Still, the canned and one-shot nature of both the underlying comments that seeded the forum and participants' own experience could have masked effects of the onboarding task on newcomer participation that would show up only in more interactive situations. (In fact, a few participants noted to us that the discussion lacked interaction between commenters).

The general future work implication is obvious—integrate into full workflows and test in the field. However, we believe the results show promise around newcomers to a discussion being able to craft prompts that according to theories of deliberation should be useful to the group.

6.8 Conclusion

In online policy discussions, people often talk past rather than with each other [42]. In this paper, we consider the role that newcomers might play in interrupting this pattern by introducing meta-talk prompts to an ongoing discussion. Meta-talk is referred to as "talk about the talk" that seeks to address points of conflict already in a discussion [207], which we relate to the type of policy discussion prompts that social studies teachers prepare for their students [177]. Drawing on the crowd-writing systems literature, we develop and evaluate key parts of a task designed to support newcomers in constructing policy discussion prompts. Our findings suggest implications for the design of crowd-writing systems, demonstrating the benefits of clear instructions and calling attention to the role that opinion and bias might play in performing judgment tasks. For online policy discussion, the findings offer considerations about the design of mechanisms that introduce newcomers to an ongoing policy discussion and existing members to new perspectives.

6.9 Acknowledgements

We thank all the Turker participants, whose thoughtful effort made this research possible. Early in the research, Poppy McLeod played an instrumental role in helping to develop the experimental design and recommending procedures for the study. During the revision cycle, we worked with Molly Feldman to more critically consider how the research findings relate to challenges in crowdsourcing and human computation. The analysis was supported by a dedicated team of research assistants: Claire Han, Anita Sharma, Simran Shinh and Varun Talwar. The research was also supported by the National Science Foundation (HCC 1314778); it was conducted while Dan Cosley was serving at the NSF and does not necessarily reflect the views of the NSF.

CHAPTER 7

GENERAL DISCUSSION

7.1 Introduction

The term “discourse architecture” is commonly used to describe how an arrangement of design features are intended to represent and elicit specific deliberative properties in an online policy discussion (see Chapter 2). The case studies present examples of fairly standard discourse architecture research. In separate experiments that vary the discussion content curation strategy (Chapter 5) and newcomer onboarding process to an online policy discussion (Chapter 6), the case studies demonstrate how manipulation of these system features affect deliberative properties of newcomer contributions, such as topic coherence.

In this way, the case study findings echo a reasonably common reprieve: “Design affects discourse.” In the practice of deliberation, decisions about the group membership, facilitation procedures, and discussion setting are designed to evoke deliberative norms of equality, respect, and reasoned opinion expression during a deliberation [21, 108]. Similarly, discussion systems, like the RegulationRoom platform, incorporate features intended to promote informed contributions that add situated knowledge to an online policy discussion (see Chapter 3). Specifically, design choices affect the deliberative properties of a policy discussion. However, the systematic review and case studies also demonstrate that applying concepts from deliberation to analyze an online policy discussion is not trivial.

A central argument in this dissertation is that the challenges involved with studying deliberativeness in online policy discussion can be addressed by tightly integrating analysis with system design research. As demonstrated through the dissertation, the process of design can lead to new insights about deliberative concepts and assumptions. Design features can also be used as a probe to investigate ways to operationalize concepts. It is also hard to separate discourse architectural decisions from analytic decisions, as design choices play into the assumptions and outcomes that are meaningful in an analysis. While design choices affect the deliberative properties of a policy discussion, I argue that the practices of design research offer tools for investigating discussion assumptions as well as deliberative concepts and their operationalization.

The following sections expand on each of these points by drawing evidence from the literature review and case studies to argue for a tight integration of analysis and the design of deliberative concepts in online policy discussion research. Contributions from the work presented in Chapters 5 and 6 are also offered as examples of how insights about online policy discussion based on concepts from deliberation may contribute back to the analysis and design of deliberation practices, whether in online, face-to-face, or even hybrid-online settings.

7.2 Design research can lead to new insights about deliberative concepts and assumptions

Deliberative concepts and assumptions about deliberation in practice are intertwined. The process of design can expose their interplay. For example, designing for newcomer contributions in online policy discussion also draws attention to the relationship between group membership and equality in other policy discussion contexts, like deliberation.

As introduced in Chapter 2, equality is an important deliberative concept that has deep roots in theory about deliberative democracy [21, 25, 88]. Assumptions about group membership play into the design choices related to equality and *vice versa*. To promote equality of access and perspective during a deliberation, participants are typically recruited and assigned so that their group membership exhibits a diversity of perspectives. Facilitation procedures during a deliberation typically treat participants as equal members of the group, with an equal opportunity to hear each other and to affect the course of their discussion [21, 144].

In practice, these design choices often translate into a group membership that is small and stable, yet restricted to a set distribution of perspectives. By designing for a small and stable group, the membership is naturally limited to people who are available, willing, and able to attend. For example, a deliberation convened in the evening may naturally exclude people who work at night, have families to tend to, or who lack access to adequate transportation. Similarly, procedures that facilitate equally among participants may not adequately account for inequalities between more and less powerful perspectives. In these ways, design choices intended to promote equality can contribute to underlying inequalities associated with deliberation [109, 192].

By designing for newcomer contributions, the case studies draw attention to issues related to the equality of newcomers in online policy discussion. By definition, newcomers arrive to a discussion with existing content to consider. Being new to a discussion that is already underway also means that newcomers have limited influence over the topics that are already under discussion. The case studies demonstrate new insights about the barriers to equality facing newcomers to an online policy discussion: (1) topic coherence is prioritized in practice, design, and analysis, (2) topic structuring decisions define the range of relevant perspectives.

7.2.1 Topic coherence is prioritized in practice, design, and analysis

Recall that if the topics in a discussion regularly shift from one to the next too quickly (i.e., low topic coherence), then the discussion is less able to deeply engage with a policy issue. During a deliberation, the topic coherence of the discussion is influenced by the facilitation procedures, which typically involve a moderator who will pose a discussion prompt and then will follow along as members of the group continue the topic by replying and responding to each other along a common line of subjects [160, 208]. For these procedures to be effective, the groups must be small enough in size that group members are able to hear each other, and stable enough that participants are able to reflect on their shared understanding of the topics [21].

Assumptions about group membership are different in an online policy discussion, as the membership will change in size and composition as participants join, leave, and return. These different assumptions about group membership require different design choices about how to promote a topically coherent discussion. For example, the moderation team at RegulationRoom devoted much of their attention to welcoming and engaging with newcomers as a strategy to promote a well-reasoned and topically coherent discussion [61] (see Chapter 3). The first case study demonstrates that system design choices, specifically related to the discussion content curation, also affect the topic coherence of newcomer contributions (Chapter 5).

Topically coherent comments are valued in online discussion. It is arguably harder to build momentum around comments that are low in topic coherence, because these comments are less likely to receive a response from the discussion [3, 104]. Topically coherent comments are also valued by analysis approaches, but comments that are less coherent are regularly dismissed as "off-topic" [94, 211] or "irrelevant" [8, 82, 142, 178, 188, 205, 212, 228]. While the coherence of an online policy discussion may depend on newcomers to continue the existing topics, design choices that prioritize topic coherence also place value on earlier topics over new topics.

These factors limit the ability of newcomers to influence the topics in an online policy discussion. Recall from Chapter 5 that curating for position agreement can yield contributions that are coherent, but the analysis also reports a nuanced relationship between a newcomers' initial position, positions prioritized in the discussion content, and the coherence of newcomer contributions to the discussion.

Rather than apply the lessons learned from the case study to design curation strategies that promote topic coherence, we might design newcomer facilitation strategies that use content curation as a tool to engage with newcomer perspectives. For example, instead of immediately introducing newcomers to the existing content at an online policy discussion, we might initially ask them to write a statement about their experience, positions, or perspectives on the issues or questions that they hope to see addressed by the discussion. As they write, text from their statement could be compared to recommend a few existing comments from the discussion for the participant to consider. A conversational assistant might present sets of comments, e.g., that are linguistically similar, express a diversity of positions, or were written by participants who have a less similar “story” about how they have been affected by the issue.

By inviting newcomers to share first, the discussion prioritizes their story over debate [192]. Several online deliberation systems include space for participants to share stories about how issues have affected them personally [45, 49, 168]. The proposed workflow offers several advancements. First, the use of text matching to identify relevant comments might facilitate a transition from personal stories [192] to collective storytelling activities (e.g., perspective taking, narrative argument) [180, 190] or into enclave deliberations that involve participants who share related stories [109].

Second, like a moderator during a deliberation, the conversational assistant may offer an opportunity for real-time reflection. Similar to “asynchronous deliberation” about judgments in crowdsourcing tasks [54], the conversational assistant may offer a newcomer an opportunity to pause and reflect on specific sets of comments, which may trigger their own internal deliberation about the issues [23, 25]. Unlike an in-person group deliberation, the personalized attention of a conversational assistant may also offer participants with a safe space to revise/refine their statements, without the social risks involved with making an off-topic comment.

7.2.2 Topic structuring decisions define the range of relevant perspectives

Where the last subsection addressed an issue of equality of access to the discussion topics, this subsection speaks to an issue about equality of perspective as the policy issues evolve past the topics structured for an online policy discussion community.

As described in Chapter 3, to prepare for an online policy discussion at RegulationRoom, the CeRI team took care to translate and triage the policy materials, so that they would be accessible to a layperson audience, but also layered the topic summaries with multiple levels of information, so that interested participants could dig deeply into a topic. Like a deliberation, the content developed for a discussion at RegulationRoom was carefully structured to respond to a time line for discussion (30-60 days) and to specific policy objectives requested by the US Federal Agency partners.

The time line and objectives for an online policy discussion community are not so specific. As a policy issue continues to evolve, its effects can touch the lives of new stakeholders, motivating new people to join in discussions about the issue online. However, it takes time to prepare an online policy discussion for new topics, perspectives, and stakeholder groups.

As a result of the time required to update the structured topics, newcomer perspectives may be viewed as off-topic given the specific state of a policy issue reflected in the discussion. New perspectives can introduce new terminology and new meaning for familiar terms. The linguistic shift associated with new perspectives can also present a barrier for older member engagement with those perspectives [39]. These factors can render newcomer perspectives at a disadvantage in an online policy discussion community.

Rather than wait for topic structure updates, an online policy discussion community might proactively identify new perspectives that deserve attention by designing for meta-talk about clarification. While the case study in Chapter 6 developed and evaluated a process for constructing policy prompts based on meta-talk about conflict, similar principles from crowd-writing systems might be applied to develop a process for facilitating meta-talk about topics and terminology to develop opportunities to hear perspectives that are perceived by existing members to be "new" to their forum.

For example, similar to the "like", "share", and "follow" buttons that accompany comments posted to a forum, we might design a flag to identify a comment that "feels interesting, but off topic." Engaging the flag would trigger a micro-task with three questions (1) what do you see as the primary topic for this thread or forum? (2) What do you hear as the primary topic in the comment? (3) Select another forum where the comment might be more useful. The newcomer, a human moderator, or a distributed moderation system might use the aggregate data collected as people perform the task to identify opportunities to clarify the topics at a forum, such as by comparing responses to tasks #1 and #2, or to identify a forum better suited for the commented perspectives, by evaluating responses to task #3. Naturally the micro-task also identifies a pool of existing members who may be interested in hearing more from the newcomer, whether at the current forum, at another forum, or at a forum newly created for topics related to their "new" perspective.

7.2.3 A dialogic moment about group membership, equality, and being new

This section has discussed the barriers to equality of access and perspective for newcomers to an online policy discussion. These barriers relate to newcomer agency over the topics under discussion as well as factors related to the topic structure that limit the influence of new perspectives. These barriers do not exist in the current practice of deliberation.

Group membership for a deliberation is often small and stable, which means that there are no newcomers to introduce (or to contend with). Shifts in the policy issue are easily managed by updating a topic structure and then recruiting a new deliberation group with a different mix of perspectives. However, inequalities underlie the practices of deliberation [108, 192]. Additionally, in their *Future Directions for Public Deliberation*, Levine et al. [133] call for deliberation procedures that scale out to include larger audiences and up to tackle policy topics that are not just local, but might be national or even global in scope. Doing so means developing procedures to integrate new topics, perspectives, and people into ongoing and possibly long term deliberations.

The topic of "Designing for newcomers in online policy discussion" may offer a useful metaphor for public deliberation researchers and practitioners as they consider ways to resolve known inequalities and to think toward the future of their field. The following are prompts based on lessons from the case studies, with specific focus on assumptions about group membership and deliberative equality. Given a set of assumptions about group membership in a deliberation (e.g., small vs large, stable vs less stable, homogeneous vs diverse perspectives):

- What barriers to entry do participants face when they are new to a deliberation group?
- What barriers does a deliberation group face when transitioning to a new topic?
- What tactics have facilitators used to adjust topic coherence (e.g., tighten, loosen)?
- What tactics have helped to introduce participants to new perspectives?
- What types of deliberative equality should participants expect?

7.3 System design features can be used as a probe to investigate operational definitions

I now turn to my second main point around the interplay of design and analysis, that design can be used to sharpen and support the crafting of operational definitions. Recall from Chapter 2 that a deliberative concept may have many operational definitions. While some view this variety as a detriment to research about policy discussion [170, 218], other deliberation scholars view variety in operational definitions as a range of tools, each with their own competitive advantage in some situations [13].

There are several common practices involved with crafting an operational definition for a deliberative concept. Stromer-Galley [207] describes a ground up approach directed by theoretical concepts that involved a ~2 month process of investigating the data from the Virtual Agora Project deliberations. Multiple researchers poured through the data and their team used statistical tools, like inter-rater reliability, to expose disagreements about how to properly observe a concept as they assembled the Measuring Deliberation's Content (MDC) analysis approach.

Chapter 3 presents another approach that draws on design choices to create an operational definition. In the example presented by the chapter, an operational definition for *Responsiveness* was developed by paying attention to the hierarchical presentation of the topic structure in the RegulationRoom discourse architecture. While operational definitions that are tightly aligned with design choices can shed light on specific circumstances, it is harder to relate bespoke definitions to more commonly referenced deliberative concepts, such as topic coherence in policy discussion.

This section presents an alternative approach that involves "sharpening" the operational definitions for a deliberative concept, by designing for concepts in system features. The case studies offer examples of two parts to this process: (1) exploring hypotheses about the operational components of a concept, and (2) probing for operational components with system design research. I argue that such a cycle can be applied to refine well-established operational definitions, like topic coherence, and to explore concepts that are less common in practice, such as meta-talk about conflict.

7.3.1 Exploring hypotheses about the operational components of a concept

System features are commonly used to achieve certain goals with respect to a deliberative property, an operational definition reflecting the concept is used to observe whether these goals are met. For instance, positioning the commenting textbox close to the prioritized comments in a discussion system increases the salience of the topics in the existing comments and can influence the content that people contribute [166] and, as demonstrated in Chapter 5, the *topic coherence* of the discussion.

By deconstructing system features of a discourse architecture into conceptual components, design research that involves the system feature can expose new insights about how to operationalize a deliberative concept. For example, the case study in Chapter 5 also reports an asymmetric relationship between a newcomers' position on a policy issue and the prioritization of positions related to the issue in the discussion thread. The case study highlights factors that may have contributed to the asymmetry, by influencing a participants perception of the "openness" and "subjectivity" expressed in the positions prioritized by each curation strategy. We might investigate these possible dimensions of topic coherence by simply modifying the content prioritized by each curation strategy through follow on research with the protocol used through the case.

By paying attention to the ways that people respond to content prioritized in an online policy discussion we might uncover other meaningful dimensions (e.g., controversy, danger, sensitivity). Future research might involve crowds or an algorithm in the process of investigating the dimensions of topic that are likely to elicit a topically coherent response. For example, the crowdsourcing platform *Flock* is a hybrid-crowd supported classifier used to classify subtle characteristics of video content, such as emotional characteristics [27]. Flock might be used to classify the "respectfulness" of a comment, which is tricky to observe because it involves evaluating the intent and effect of a comment [211]. Unlike the practice of crafting operational definitions by pouring over policy discussion transcripts, design research with system features can be used to "prototype" operational definitions related to a deliberative concept.

7.3.2 Probing for operational components with system design research

Beyond careful observation of how system features are used and whether their intended goals are achieved, discourse architectures can also be explicitly designed to support research around operational definitions. In the theory-practice debate among deliberation scholars, operational definitions might be a middle ground. Theorists can point at how a theory has been implemented improperly by an operational definition and practitioners can use the impracticality of some operational definitions to point to policy discussion assumptions that are unaccounted for by theory. However, it may be difficult to facilitate a reasoned deliberation about how to operationalize a theoretical concept when there is little evidence as to its existence in practice.

Designing for a concept may be a way to probe for possible operational components of a rarely observed deliberative concept. For example, the deliberative concept *meta-talk* is rare in an online policy discussion. Furthermore, observations of meta-talk in online policy discussion are commonly related to the level of incivility in the discussion, rather than opportunities for consensus, clarification, or for conflict (see Chapter 2). The case study in Chapter 6 develops and evaluates a system feature that was modeled after *meta-talk about conflict*—i.e., a process for newcomers to craft policy prompts by synthesizing two existing comments.

One reason a concept might be rare is that the operational definition is not complete or explicit enough to explain the concept. In designing for meta-talk about conflict, I found that the MDC operational definition was not detailed enough to be *designable*. The following statement reflects the MDC operational definition for meta-talk about conflict: "highlighting some disagreement or conflict in the group," as indicated by statements like, "I sense some disagreement around [...]" [207, pg. 25]. This statement is hard to design with because it leaves critical questions open for interpretation, such as how to present a prior disagreement in a group and then request that the group revisit (or resolve) the disagreement.

To identify a designable definition of meta-talk about conflict, I surveyed practitioner focused literature about discussion facilitation, knowing that moderators tend to raise meta-talk during a policy discussion [207]. The case study presented in Chapter 6, considers a more nuanced operational definition of meta-talk that includes three components [193]: i.e., a reference, logical operators to evaluate the reference, and a verb that requests an action related to the reference, such as clarify, tell, or argue. This definition resolved many of the questions about meta-talk that I felt were left open in MDC.

Through the process of designing for meta-talk, I realized that standard commenting systems may not be conducive for meta-talk. Most commenting systems also direct comment responses to a specific comment and commenter, which may make it difficult for people to highlight evidence of a disagreement across multiple comments and for select groups of people to consider. The system feature developed for the case study presents a pair of comments for participants to engage with, where standard commenting systems display existing comments in a list or thread. Rather than focus on standard commenting systems, we might consider investigating meta-talk at systems that allow for comment collection or synthesis (e.g., *ConsiderIt* [118], *Cohere* [200], *Wikum* [230]).

7.3.3 A dialogic moment to sharpen operational definitions with design research

The central argument of this section is that system design can be used as a resource to investigate ways to operationalize a deliberative concept within a discourse architecture. Through a practice of designing for a deliberative concept in different contexts, researchers might develop a broad range of operational tools for observing the concept, each with a competitive advantage in some circumstances [13]. For example, the operational definitions for topic coherence might include slight modifications for an online policy discussion with threaded responses [205], unthreaded discussion [204], synchronous discussion [211], discussion curated by distributed moderation [33], or discussion within hashtags [70].

The practice of system design for a deliberative concept may also offer opportunities for theorists and system designers to collaborate by "prototyping" concepts as system features at an online policy discussion. As recounted in the section, the process of designing for meta-talk about conflict revealed aspects of the MDC definition that were challenging to design for and ultimately led my design process to a related, but more concrete definition. By developing and then evaluating a system feature intended to craft policy discussion prompts in the style of meta-talk about conflict, I realized that some aspects of standard commenting systems may also limit the occurrence of meta-talk in online policy discussion. A similar design process might be pursued with other deliberative concepts to similarly sharpen their operational definition(s).

A residual benefit to such a design-driven inquiry into deliberative concepts is that it might encourage dialogue among system designers and theorists who study policy discussion in different contexts. We might consider facilitating a weekend-long dialogue about the theory-practice gap with members of the deliberation community. On the first day, we might assign participants to small teams of three—i.e., one theorist, empiricist, and practitioner—and challenge each team to propose deliberation facilitation procedures to reflect a set of deliberative concepts. During the second day, we might use the work developed during the first day to conduct mock deliberations and test out evaluation strategies. This process of developing, enacting, and evaluating the proposed procedures might refine and generate new concepts, but it may also elicit useful dialogue about how theorists, empiricists, and practitioners draw meaning from and make use of deliberative concepts.

7.4 Architectural decisions, research decisions, and discussion outcomes are hard to separate

Each of the prior sections argue for a tight integration in the analysis and design of deliberative concepts to advance online policy discussion research and system design. Designing for a deliberative concept can expose underlying relationships with assumptions about policy discussion. The process of designing for a deliberative concept can also draw attention to new operational dimensions and strategies for observing the concept at an online policy discussion. Observations from online policy discussion can inspire new deliberative concepts, in the same way that deliberation scholars incorporate observations from the practice of deliberation into an analysis.

With so much emphasis on analysis concepts and system design choices, it is not hard to lose track of the fact that policy discussion's have outcomes beyond research. This reality plays an important part in widening the deliberation theory-practice gap, as deliberation protocols are typically modified to meet the needs of specific stakeholder groups, policy issues, and social contexts [108]. These considerations play into assumptions about group membership, facilitation procedures, and the settings for a policy discussion.

7.4.1 Some perspectives on a policy issue are easy to access

Conducting online policy discussion research with Amazon Mechanical Turk (AMT) crowd workers (called "Turkers") is convenient. As an online crowd labor market, the design of AMT prioritizes low-latency and programmatic access to human intelligence. To recruit participants for each case study, I posted a human intelligence task (HIT) to the AMT marketplace. However, this recruitment procedure intentionally limited the participation of other stakeholders, such as the people who hire Turkers (called "Requesters"), manage the labor market, and oversee the participation agreement. As AMT has been used to target content at other platforms, such as by hiring Turkers to post fake product reviews at other websites and to artificially inflate the popularity of content at online discussion forums, a formal deliberation about issues related to crowd labor at AMT may also involve voices from platforms that have been adversely affected by crowdsourced mischief [103, 233].

Demographically, the Turker community also includes people from around the world and from a wide range of social, educational, and economic status [102]. The decision to present the policy context and discussion in the English language may have limited participation from people whose first language is not English. Similarly, the recruitment strategy priced the task on an estimate of the hourly minimum wage for US workers based in the state of New York. While higher than the US Federal minimum wage as well as many other states in the US, the New York minimum wage is lower than the state minimum wage in Washington, where Amazon is incorporated. Additionally, the minimum wage in New York is less than the minimum wage in many countries (i.e., Luxembourg, Netherlands, Ireland, France, Germany). Therefore, the recruitment strategy for the case studies may have favored a subset of US-based Turkers, even though the HIT was available to all interested.

7.4.2 Research objectives can be at odds with policy outcomes

The case studies consider a policy context that prioritizes the Turker perspective of the AMT participation agreement over other perspectives. In the process of developing this material, I personally worked with leaders in the Turker community to confirm that the policy summary spoke to issues that matter to Turkers. In so doing, the policy context applied through the case study research highlights just one among several perspectives about crowd labor policies at AMT. While these decisions were meant to interest Turkers in the policy discussion, other decisions related to the topics and positions served research objectives.

The specific policy topic, "what should happen when a HIT is rejected?" surfaced as the most discussed topic in prior research [153], though the prior research solicited comments from Turkers about a range of other AMT platform policy issues, such as paying taxes, Amazon's hands-off approach to AMT, and concerns related to personal privacy at AMT. Furthermore, this policy topic was interesting for research purposes, because Turkers have different and conflicting opinions about how to address the issue. As Turkers are consistently divided about "partial payment" as a viable solution, I was able to conduct a series of experiments varying whether opposing perspectives were prioritized in the existing thread of comments or as part of a crowd-writing task (Chapters 5 & 6).

Seeding the discussion with an equal number of comments in support and opposition also elevated the "partial payment" debate. Alternatively the case studies might have focused on a more popular proposal, such as a "second chance" to correct mistakes; however, the controversy surrounding partial payment enabled research about how newcomers respond when presented with policy positions that are opposed to their own. The case studies therefore prioritize research objectives in the selection of policy topics, rather than the popularity of policy options.

7.4.3 There are ethical considerations when organizing a discussion within some settings

While AMT does not provide Turkers a way to communicate with each other, or even for Requesters to communicate with Turkers who they have not hired recently (within 45 days), there are numerous online discussion forums related to Turking at AMT. An entire ecosystem of online discussion has emerged around Turker relationships with AMT, Requesters, and to each other [103, 147, 191]. Several forums are used to share tips about specific HITs and about Requesters (e.g., *HITsWorthTurkingFor*, *Turkopticon*, *Crowd-workers.com*). This ecosystem of online discussion also includes several sites where Turkers do meet and discuss political and policy issues, often about AMT, but not always (e.g., *TurkerNation*, *MTurkGrind*). Additionally, the ecosystem includes the *Dynamo* platform, which offers a deliberation space for online activism, with tools for Turkers to build ideas and support for political campaigns related to AMT. Rather than recruit Turkers to an online discussion through a HIT on AMT, the case study research might have focused on discussions at sites in-the-wild that are regularly frequented by Turkers.

There are ethical considerations related to using a HIT posted to AMT to recruit participants to discuss issues about HIT rejection and the AMT participation agreement. While the recruited Turkers were participants in a series of research studies, they were also temporary employees managed by the research team. This meant that the research team had the authority and resources through AMT to reject the services of any Turker participant. We did not. However, this power asymmetry speaks to the different hats that experimentalists wear in online policy discussion research. In some contexts, our role is to educate and inform, by crafting policy materials and questions intended to spark interest. In other cases, we think strategically about how phrasing in the policy materials or the selection of content respond to research objectives. When working with AMT and within other online labor markets, we are also managers, responsible to our employees, temporary as they may be [151].

7.4.4 Some discourse architectural decisions are worth policy discussion in their own right

Deliberation organizers pay careful attention to the policy content included in a deliberation. As discussed in Chapter 2, the background materials, instructions, and participant training activities associated with a policy deliberation are all carefully vetted, so as to present the issues for deliberation in ways that are accessible for the participants [108]. Similar to the steps taken by deliberation organizers, the CeRI team devoted a substantial amount of time to prepare policy materials for a discussion at RegulationRoom (see Chapter 3). While these preparatory steps are used to structure the topics that organizers expect (or hope) that the discussion participants will address, human-moderators typically manage the participant engagement with topics, perspectives, and each other during a policy discussion [145, 160].

In online policy discussion, system design choices related to the curation of content play a central role in the deliberativeness of a policy discussion. In Chapter 4, decisions about what positions of an issue to prioritize in the discussion content factored into the coherence of new contributions to the topics already under discussion. In Chapter 5, participants in a crowd-writing process to craft a policy discussion prompt were more likely to produce an acceptable prompt when asked in the write condition to synthesize two existing comments that were further apart than similar in their position on an issue. These and other participant-level observations from the case studies highlight the role that content curation decisions play into the dynamics of a policy discussion, much like a moderator during a deliberation.

Design choices and practices related to content curation at online discussion systems deserve policy discussion in their own right. "Freedom of Speech" is commonly referenced as a guide for decisions about how to curate online discussion platforms [96]. The First Amendment to the US Constitution states that "Congress shall make no law [...] abridging freedom of speech." However, the First Amendment protects people from censorship by the US government, it does not protect against censorship at online platforms managed by private companies. Furthermore, people may be permitted to contribute content to an online discussion system, but computing systems play a pronounced role in determining what content other people are presented, when it is presented, and in what order.

A slightly more nuanced content curation challenge is how much perspective diversity should people be exposed to during an online policy discussion? People prefer different levels of "disagreement" during a deliberation [209] and different amounts of "perspective diversity" in their online consumption of policy information [166]. People may use content filtering features in an online discussion system to selectively avoid topics and perspectives that are opposed to their own [176]. Similarly, personalized recommendation systems in content curation can automatically nudge people toward content that reinforces their existing preferences [80]. Rather than focus on what types of content individual participants prefer, deliberation organizers consider the various outcomes that specific assumptions about the discussion group might yield (e.g., membership, facilitation procedures, setting) [107, 108].

Another challenge related to content curation is how to manage the equality of access and perspective during a policy discussion. As widely acknowledged by deliberation scholars, facilitation procedures that enable an equal opportunity to speak can yield inequalities between participants with more and less time, training, and influence [109, 192]. Similar concerns play out in decisions about how content is curated at an online discussion system. For example, the recent (and regular) bot-powered attacks against discussion at online systems demonstrate how content curation mechanisms can be exploited to distribute propaganda and misinformation [35]. Rather than focus on how to block specific actors, deliberation organizers consider ways to prepare and support groups of participants through their process of deliberation [109, 105] and dialogue about alternate perspectives of an issue [10].

There is also a human cost associated with the process of training an automated content curation system. Repeated exposure to the graphic content that online systems try to keep off of their platforms can be psychologically damaging for the human content moderators involved [182]. Although deliberation is often referred to as a public engagement process that can yield well-reasoned decisions [207], deliberation is also used to foster mutual understanding and community through dialogue [10, 63, 190]. A regular process of deliberation related to the work of content moderation may offer useful insights for the people who manage and develop policies that govern online platforms, but may also offer some relief, community, and power for the content moderators involved in this work.

As content curation decisions determine the topics and perspectives to prioritize in an online policy discussion, these system design choices affect the allocation of power in a discussion. In the practice of deliberation, organizers pay careful attention to how design choices related to the group membership, facilitation procedures, and discussion setting for a deliberation promote equality in the discussion, legitimacy in the outcome, and limit the influence of powerful interests [108]. My hope is that this examination of the use and usefulness of deliberative concepts will help policy makers, system designers, and others to develop and evaluate power in online policy discussion.

In short, system design choices are political.

BIBLIOGRAPHY

[1] Tanja Aitamurto and Helene Landemore. Crowdsourced Deliberation: The Case of the Law on Off-Road Traffic in Finland. *Policy & Internet*, 8(2): 174–196, 2016.

[2] Augusta Isabella Alberici and Patrizia Milesi. Online discussion, politicized identity, and collective action. *Group Processes & Intergroup Relations*, 19(1):43–59, 1 2016. doi: 10.1177/1368430215581430.

[3] Jaime Arguello, Brian S. Butler, Elisabeth Joyce, Robert Kraut, Kimberly S. Ling, Carolyn Rosé, and Xiaoqing Wang. Talk to me: foundations for successful individual-group interactions in online communities. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 959–968, 2006.

[4] Blake E. Ashforth, David M. Sluss, and Alan M. Saks. Socialization tactics, proactive behavior, and newcomer learning: Integrating socialization models. *Journal of Vocational Behavior*, 70(3):447–462, 2007. URL `https://doi.org/10.1016/j.jvb.2007.02.001`.

[5] Lars Backstrom, Jon Kleinberg, Lillian Lee, and Cristian Danescu-Niculescu-Mizil. Characterizing and curating conversation threads: expansion, focus, volume, re-entry. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 13–22, New York, NY, 2013. ACM. doi: 10.1145/2433396.2433401. URL `http://dl.acm.org/citation.cfm?id=2433401`.

[6] Martina Balestra, Orit Shaer, Johanna Okerlund, Madeleine Ball, and Oded Nov. The effect of exposure to social annotation on online informed

consent beliefs and behavior. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 900–912, San Francisco, 2016. ACM.

[7] Albert Bandura. *Self-efficacy: The exercise of control*. W. H. Freeman and Company, New York, NY, 1997.

[8] Janne Berg. The impact of anonymity and issue controversiality on the quality of online discussion. *Journal of Information Technology & Politics*, 13 (1):37–51, 1 2016. doi: 10.1080/19331681.2015.1131654.

[9] Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. Soylent: a word processor with a crowd inside. *Communications of the ACM*, 58(8):85–94, 2015.

[10] Laura W. Black. Deliberation, Storytelling, and Dialogic Moments. *Communication Theory*, 18(1):93–116, 1 2008. doi: 10.1111/j.1468-2885.2007.00315.x.

[11] Laura W. Black. Blog, chat, edit, text, or tweet? Using online tools to advance adult civic engagement. *New Directions for Adult and Continuing Education*, 2012(135):71–79, 2012.

[12] Laura W Black. Framing democracy and conflict through storytelling in deliberative groups. *Journal of Public Deliberation*, 9(1), 2013.

[13] Laura W. Black, Stephanie Burkhalter, John Gastil, and Jennifer Stromer-Galley. Methods for analyzing and measuring group deliberation. *Sourcebook of political communication research: Methods, measures, and analytical techniques*, pages 323–345, 2010.

[14] Laura W. Black, Howard T. Welser, Dan Cosley, and Jocelyn M. DeGroot. Self-Governance Through Group Discussion in Wikipedia. *Small Group Research*, 42(5):595–634, 10 2011. doi: 10.1177/1046496411406137.

[15] Robin Blom, Serena Carpenter, Brian J. Bowe, and Ryan Lange. Frequent Contributors Within U.S. Newspaper Comment Forums. *American Behavioral Scientist*, 58(10):1314–1328, 9 2014. doi: 10.1177/0002764214527094.

[16] danah boyd. Why youth (heart) social network sites: The role of networked publics in teenage social life. *MacArthur foundation series on digital learning–Youth, identity, and digital media volume*, pages 119–142, 2007.

[17] Jan Brett. *Goldilocks and the three bears*. Penguin, 2016.

[18] Rupert Brown. *Group processes: Dynamics within and between groups.* Basil Blackwell, 1988.

[19] Moira Burke, Cameron Marlow, and Thomas Lento. Feed me: motivating newcomer contribution in social network sites. *Proceedings of the SIGCHI conference on*, 2009. URL http://dl.acm.org/citation.cfm?id=1518847.

[20] Moira Burke, Robert E. Kraut, and Elisabeth Joyce. Membership Claims and Requests: Conversation-Level Newcomer Socialization Strategies in Online Groups. *Small Group Research*, 41(1):4–40, 2 2010. doi: 10.1177/1046496409351936. URL http://sgr.sagepub.com/cgi/doi/10.1177/1046496409351936.

[21] Stephanie Burkhalter, John Gastil, and Todd Kelshaw. A conceptual definition and theoretical model of public deliberation in small face-to-face groups. *Communication Theory*, 12(4):398–422, 2002.

[22] Bill Buxton. *Sketching user experiences: getting the design right and the right design*. Morgan Kaufmann, 2010.

[23] Joseph N. Cappella, Vincent Price, and Lilach Nir. Argument Repertoire as a Reliable and Valid Measure of Opinion Quality: Electronic Dialogue During Campaign 2000. *Political Communication*, 19(1):73–93, 1 2002. doi: 10.1080/105846002317246498.

[24] Alissa Centivany and Bobby Glushko. 'Popcorn Tastes Good': Participatory Policymaking and Reddit's' Amageddon'. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI16), San Jose, CA*, 2016.

[25] Simone Chambers. Deliberative democratic theory. *Annual review of political science*, 6(1):307–326, 2003.

[26] Gilad Chen, Stanley M. Gully, and Dov Eden. Validation of a new general self-efficacy scale. *Organizational research*, 2001. URL http://journals.sagepub.com/doi/abs/10.1177/109442810141004.

[27] Justin Cheng and Michael S Bernstein. Flock: Hybrid crowd-machine learning classifiers. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 600–611, 2015.

[28] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. Antisocial behavior in online discussion communities. *arXiv preprint arXiv:1504.00680*, 2015.

[29] Boreum Choi, Kira Alexander, Robert E. Kraut, and John M. Levine. Socialization tactics in wikipedia and their effects. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work - CSCW '10*, page

107, New York, New York, USA, 2010. ACM Press. ISBN 9781605587950. doi: 10.1145/1718918.1718940.

[30] Hoon-Seok Choi and Leigh Thompson. Old wine in a new bottle: Impact of membership change on group creativity. *Organizational Behavior and human decision*, 98(2):121–132, 2005. URL https://doi.org/10.1016/j.obhdp.2005.06.003.

[31] Christopher H Clark, Daniel T Bordwell, and Patricia G Avery. Gender and Public Issues Deliberations in Named and Anonymous Online Environments. *Journal of Public Deliberation*, 11(2), 2015.

[32] Herbert H. Clark and Susan E. Brennan. Grounding in communication. *Perspectives on socially shared cognition*, 13(1):127–149, 1991.

[33] Kevin Coe, Kate Kenski, and Stephen A. Rains. Online and Uncivil? Patterns and Determinants of Incivility in Newspaper Website Comments. *Journal of Communication*, 64(4):658–679, 8 2014. doi: 10.1111/jcom.12104.

[34] Stephen Coleman and John Gøtze. *Bowling together: Online public engagement in policy deliberation*. Hansard Society London, 2001.

[35] Nicholas Confessore, Gabriel J.X. Dance, Richard Harris, and Mark Hansen. The Follower Factory, 1 2018. URL https://www.nytimes.com/interactive/2018/01/27/technology/social-

[36] Jeff Conklin and Michael L. Begeman. gIBIS: a hypertext tool for exploratory policy discussion. *ACM Transactions on Information Systems*, 6 (4):303–331, 10 1988. doi: 10.1145/58566.59297.

[37] Maria J D'Agostino, Richard W Schwester, and Marc Holzer. Enhancing

the prospect for deliberative democracy: The AmericaSpeaks model. *The Innovation Journal: The Public Sector Innovation Journal*, 11(1), 2006.

[38] Peng Dai and Daniel Sabey Weld. Decision-Theoretic Control of Crowd-Sourced Workflows. *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 7 2010. URL `https://www.aaai.org/ocs/index.php/AAAI/AAAI10/paper/viewPaper/18`

[39] Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. No country for old members: User life-cycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*, pages 307–318, 2013.

[40] Todd Davies and Reid Chandler. Online Deilberation Design: Choices, Criteria, and Evidence. *Democracy in motion: Evaluating the practice and impact of deliberative civic engagement*, pages 103–131, 2011.

[41] Todd Davies and Seeta Peña Gangadharan. Online deliberation: Design, research, and practice. 2009.

[42] Richard Davis. *The web of politics: The Internet's impact on the American political system*. Oxford University Press, 1999.

[43] Fiorella De Cindio and Cristian Peraboni. Design issues for building deliberative digital habitats. *Online Deliberation*, page 41, 2010.

[44] Fiorella De Cindio and Douglas Schuler. Beyond community networks: From local to global, from participation to deliberation. *The journal of community informatics*, 8(3), 2012.

[45] Fiorella De Cindio, Cristian Peraboni, and Leonardo Sonnante. A two-room e-deliberation environment. *Proceedings of Tools for Participation: Collaboration, deliberation, and decision support*, pages 47–59, 2008.

[46] Anna De Liddo, Ágnes Sándor, and Simon Buckingham Shum. Contested Collective Intelligence: Rationale, Technologies, and a Human-Machine Annotation Study. *Computer Supported Cooperative Work (CSCW)*, 21(4-5): 417–448, 10 2012. doi: 10.1007/s10606-011-9155-x.

[47] John Dewey, Jo Ann Boydston, and Abraham Edel. *The later works, 1925-1953. 7. 1932:[Ethics]*, volume 7. SIU Press, 2008.

[48] Nicholas Diakopoulos and Mor Naaman. Topicality, time, and sentiment in online news comments. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pages 1405–1410, 2011.

[49] Jill P Dimond, Michaelanne Dye, Daphne LaRose, and Amy S Bruckman. Hollaback!: the role of storytelling online in a social movement organization. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 477–490, 2013.

[50] Carl DiSalvo. Design and the Construction of Publics. *Design issues*, 25(1): 48–63, 2009.

[51] Steven P. Dow, Alana Glassco, Jonathan Kass, Melissa Schwarz, Daniel Schwartz, and Scott Klemmer. Parallel prototyping leads to better design results, more divergence, and increased self-efficacy. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 17(4), 2010. URL `http://dl.acm.org/citation.cfm?id=1879836`.

[52] Steven P. Dow, Julie Fortuna, Dan Schwartz, and Beth Altringer. Prototyping dynamics: sharing multiple designs improves exploration, group rapport, and results. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2807–2816, 2011. URL `http://dl.acm.org/citation.cfm?id=1979359`.

[53] Steven P. Dow, An Kulkarni, Scott Klemmer, and Björn Hartmann. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 1013–1022, 2012.

[54] Ryan Drapeau, Lydia B. Chilton, Jonathan Bragg, and Daniel S. Weld. MicroTalk: Using Argumentation to Improve Crowdsourcing Accuracy. *Human Computation and Crowdsourcing*, 2016.

[55] John S Dryzek. Deliberative Democracy in Divided Societies: Alternatives to Agonism and Analgesia. *Source: Political Theory*, 33(2):218–242, 2005.

[56] Geoffrey B. Duggan and Stephen J. Payne. Skim reading by satisficing: evidence from eye tracking. *Proceedings of the SIGCHI Conference on*, 2011. URL `http://dl.acm.org/citation.cfm?id=1979114`.

[57] David Dutwin. The character of deliberation: Equality, argument, and the formation of public opinion. *Journal of Public Opinion Research*, 15(3): 239–264, 2003.

[58] Steve M. Easterbrook, Eevi E. Beck, James S. Goodlet, Lydia Plowman, Mike Sharples, and Charles C. Wood. A survey of empirical studies of conflict. In *CSCW: Cooperation or Conflict?*, pages 1–68. Springer, 1993.

[59] Andrew J. Elliot and Patricia G. Devine. On the motivational nature of cognitive dissonance: Dissonance as psychological discom-

fort. *Journal of personality and social psychology*, 67(3):382, 1994. URL `http://psycnet.apa.org/journals/psp/67/3/382/`.

[60] Dima Epstein, Cynthia Farina, and Josiah Heidt. The value of words: Narrative as evidence in policy making. *Evidence & Policy: A Journal of Research, Debate, and Practice*, 10(2):243–258, 2014.

[61] Dmitry Epstein and Gilly Leshed. The Magic Sauce: Practices of Facilitation in Online Policy Deliberation. *Journal of Public Deliberation*, 12(1), 2016.

[62] Tobias Escher, Dennis Friess, Katharina Esau, Jost Sieweke, Ulf Tranow, Simon Dischner, Philipp Hagemeister, and Martin Mauve. Online Deliberation in Academia: Evaluating the Quality and Legitimacy of Cooperatively Developed University Regulations. *Policy & Internet*, 9(1):133–164, 3 2017. doi: 10.1002/poi3.119.

[63] Oliver Escobar. The dialogic turn: Dialogue for deliberation. *In-Spire Journal of Law, Politics and Societies*, 2009.

[64] Alek Felstiner. Working the crowd: employment and labor law in the crowdsourcing industry. *Berkeley Journal of Employment and Labor Law*, pages 143–203, 2011.

[65] Leon Festinger. *A theory of cognitive dissonance*, volume 2. Stanford university press, 1962.

[66] Casey Fiesler, Shannon Morrison, R. Benjamin Shapiro, and Amy S. Bruckman. Growing Their Own. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW*

'17, pages 1375–1386, New York, New York, USA, 2017. ACM Press. ISBN 9781450343350. doi: 10.1145/2998181.2998210.

[67] James S. Fishkin. *Democracy and deliberation: New directions for democratic reform*. 1991.

[68] James S. Fishkin. *When the people speak: Deliberative democracy and public consultation*. 2011.

[69] Rolf Fredheim, Alfred Moore, and John Naughton. Anonymity and Online Commenting: The Broken Windows Effect and the End of Drive-by Commenting. In *Proceedings of the ACM Web Science Conference*, page 11, Oxford, United Kingdom, 2015. ACM. doi: 10.1145/2786451.2786459.

[70] Deen Freelon. Discourse architecture, ideology, and democratic norms in online political discussion. *New Media & Society*, 17(5):772–791, 2015.

[71] Dennis Friess and Christiane Eilders. A Systematic Review of Online Deliberation Research. *Policy & Internet*, 7(3):319–339, 9 2015. ISSN 19442866. doi: 10.1002/poi3.95. URL http://doi.wiley.com/10.1002/poi3.95.

[72] Susan R Fussell and Leslie D Setlock. Computer-mediated communication. *The Oxford Handbook of Language and Social Psychology*, page 471, 2014.

[73] James Gastil. *By popular demand: Revitalizing representative democracy through deliberative elections*. University of California Press, 2000.

[74] John Gastil. Is Face-to-Face Citizen Deliberation a Luxury or a Necessity? *Political Communication*, 17(4):357–361, 10 2000. doi: 10.1080/10584600050178960.

[75] John Gastil. *Political communication and deliberation*. Sage, 2008.

[76] John Gastil and Laura Black. Public deliberation as the organizing principle of political communication research. *Journal of Public Deliberation*, 4 (1), 2008.

[77] John Gastil and James P. Dillard. The aims, methods, and effects of deliberative civic education through the National Issues Forums. *Communication education*, 1999.

[78] John Gastil and Peter Levine. *The deliberative democracy handbook: Strategies for effective civic engagement in the twenty-first century*. Jossey-Bass San Francisco, 2005.

[79] John Gastil, Laura Black, and Kara Moscovitz. Ideology, Attitude Change, and Deliberation in Small Face-to-Face Groups. *Political Communication*, 25(1):23–46, 2 2008. doi: 10.1080/10584600701807836.

[80] Tarleton Gillespie. The Relevance of Algorithms. *Media technologies: Essays on communication, materiality, and society*, page 167, 2014.

[81] Erving Goffman. Replies and responses. *Language in society*, 5(03):257–313, 1976.

[82] Patricia Gonçalves da Conceição Rossini and Vanessa Veiga de Oliveira. International Journal of Communication. *International Journal of Communication*, 10(1), 2016.

[83] Eric Gordon and Edith Manosevitch. Augmented deliberation: Merging physical and virtual interaction to engage communities in urban planning. *New Media & Society*, 13(1):75–95, 2 2011. doi: 10.1177/1461444810365315.

[84] Todd Graham. Beyond "Political" Communicative Spaces: Talking Politics on the Wife Swap Discussion Forum. *Journal of Information Technology & Politics*, 9(1):31–45, 1 2012. doi: 10.1080/19331681.2012.635961.

[85] Todd Graham and Scott Wright. Discursive Equality and Everyday Talk Online: The Impact of "Superparticipants". *Journal of Computer-Mediated Communication*, 19(3):625–642, 4 2014. doi: 10.1111/jcc4.12016.

[86] Ali Gürkan, Luca Iandoli, Mark Klein, and Giuseppe Zollo. Mediating debate through on-line large-scale argumentation: Evidence from the field. *Information Sciences*, 180(19):3686–3702, 2010. doi: 10.1016/j.ins.2010.06.011.

[87] Jürgen Habermas. Strukturwandel der öffentlichkeit. 1962.

[88] Jürgen Habermas. *The theory of communicative action*, volume 2. Beacon press, 1985.

[89] Jürgen Habermas. Three normative models of democracy. *Constellations*, 1(1):1–10, 1994.

[90] Carole Hahn. *Becoming political: Comparative perspectives on citizenship education*. Suny Press, 1998.

[91] Nathan Hahn, Joseph Chang, Ji Eun Kim, and Aniket Kittur. The Knowledge Accelerator: Big Picture Thinking in Small Pieces. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2258–2270, 2016.

[92] Maarten Hajer. Policy without polity? Policy analysis and the institutional void. *Policy sciences*, 36(2):175–195, 2003.

[93] Aaron Halfaker, Oliver Keyes, and Dario Taraborelli. Making peripheral participation legitimate. In *Proceedings of the 2013 conference on Computer supported cooperative work - CSCW '13*, page 849, New York, New York, USA, 2013. ACM Press. ISBN 9781450313315. doi: 10.1145/2441776.2441872. URL `http://dl.acm.org/citation.cfm?doid=2441776.2441872`.

[94] Daniel Halpern and Jennifer Gibbs. Social media as a catalyst for online deliberation? Exploring the affordances of Facebook and YouTube for political expression. *Computers in Human Behavior*, 29(3):1159–1168, 2013. doi: 10.1016/j.chb.2012.10.008.

[95] Roderick Hart and Sharon Jarvis. We the people: The contours of lay political discourse. In *The National Issues Convention experiment in political communication*, pages 59–84. 1999.

[96] Miguel Helft. Facebook wrestles with free speech and civility, 12 2010.

[97] Diana E. Hess. *Controversy in the classroom: The democratic power of discussion*. Routledge, 2009.

[98] Diana E. Hess and Paula McAvoy. *The political classroom: Evidence and ethics in democratic education*. Routledge, 2014.

[99] Lorraine Higgins, Elenore Long, and Linda Flower. Community literacy: A rhetorical model for personal and public inquiry. *Community Literacy Journal*, 1(1):9, 2006.

[100] Gary Hsieh, Youyang Hou, Ian Chen, and Khai N. Truong. Welcome!: Social and psychological predictors of volunteer socializers in online com-

munities. In *Computer Supported Cooperative Work*, pages 827–838, 2013. URL `http://dl.acm.org/citation.cfm?id=2441870`.

[101] Ronald J. Hustedde. An evaluation of the National Issues Forum methodology for stimulating deliberation in rural Kentucky. *Community Development*, 27(2):197–210, 1996.

[102] Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 64–67, 2010.

[103] Lilly C. Irani and M. Silberman. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 611–620. ACM, 2013.

[104] Elisabeth Joyce and Robert E. Kraut. Predicting continued participation in newsgroups. *Journal of Computer-Mediated Communication*, 11(3):723–747, 2006.

[105] Christopher F. Karpowitz and Jane Mansbridge. Disagreement and consensus: The need for dynamic updating in public deliberation. *Journal of Public deliberation*, 1(1), 2005.

[106] Christopher F. Karpowitz and Tali Mendelberg. Groups and Deliberation. *Swiss Political Science Review*, 13(4):645–662, 12 2007. doi: 10.1002/j.1662-6370.2007.tb00092.x.

[107] Christopher F. Karpowitz and Tali Mendelberg. An experimental approach to citizen deliberation. In *Cambridge handbook of experimental political science*, pages 258–272. 2011.

[108] Christopher F. Karpowitz and Chad Raphael. *Deliberation, democracy, and civic forums: Improving equality and publicity*. Cambridge University Press, 2014.

[109] Christopher F. Karpowitz, Chad Raphael, and Allen S. Hammond. Deliberative Democracy and Inequality: Two Cheers for Enclave Deliberation among the Disempowered. *Politics & Society*, 37(4):576–615, 12 2009. doi: 10.1177/0032329209349226.

[110] Joy Kim and Andres Monroy-Hernandez. Storia: Summarizing social media content based on narrative theory using crowdsourcing. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 1018–1027, 2016.

[111] Joy Kim, Sarah Sterman, Allegra Argent Beal Cohen, and Michael S. Bernstein. Mechanical Novel: Crowdsourcing Complex Work through Reflection and Revision. In *Computer Supported Cooperative Work and Social Computing*, pages 233–245. ACM, 2016. URL `https://doi.org/10.1145/2998181.2998196`.

[112] Nam Wook Kim, Jonghyuk Jung, Eun-Young Ko, Songyi Han, Chang Won Lee, Juho Kim, and Jihee Kim. Budgetmap: Engaging taxpayers in the issue-driven classification of a government budget. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 1028–1039, 2016.

[113] Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E Kraut. Crowdforge: Crowdsourcing complex work. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 43–52, 2011.

[114] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1301–1318, 2013.

[115] Mark Klein. How to harvest collective wisdom on complex problems: An introduction to the mit deliberatorium. *Center for Collective Intelligence working paper*, 2011.

[116] Robert E. Kraut, Moira Burke, John Riedl, and Paul Resnick. Dealing with Newcomers. *Evidence-based Social Design: Mining the Social Sciences to Build Online Communities*, 1(1):42, 2010.

[117] Robert E. Kraut, Paul Resnick, Sara Kiesler, Moira Burke, Yan Chen, Niki Kittur, Joseph Konstan, Yuqing Ren, and John Riedl. *Building successful online communities: Evidence-based social design*. Mit Press, 2012.

[118] Travis Kriplean, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett. Supporting reflective public thought with considerit. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 265–274, 2012.

[119] Travis Kriplean, Michael Toomim, Jonathan Morgan, Alan Borning, and Andrew Ko. Is this what you meant?: promoting listening on the web with reflect. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1559–1568, 2012.

[120] Travis Kriplean, Caitlin Bonnar, Alan Borning, Bo Kinney, and Brian Gill. Integrating on-demand fact-checking with public dialogue. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social*

computing - CSCW '14, pages 1188–1199, New York, New York, USA, 2014. ACM Press. ISBN 9781450325400. doi: 10.1145/2531602.2531677. URL http://dl.acm.org/citation.cfm?doid=2531602.2531677.

[121] Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, and Scott R. Klemmer. Peer and self assessment in massive online classes. *ACM Transactions on Computer-Human Interaction*, 20(6): 1–31, 12 2013. ISSN 10730516. doi: 10.1145/2505057. URL http://dl.acm.org/citation.cfm?doid=2562181.2505057.

[122] Werner Kunz and Horst Rittel. Issues as elements of information systems. 1970.

[123] Cliff Lampe and Paul Resnick. Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 543–550, 2004.

[124] Cliff Lampe, Paul Zube, Jusil Lee, Chul Hyun Park, and Erik Johnston. Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums. *Government Information Quarterly*, 31(2):317–326, 4 2014. doi: 10.1016/j.giq.2013.11.005.

[125] Anders Olof Larsson. Everyday elites, citizens, or extremists? Assessing the use and users of non-election political hashtags. *MedieKultur: Journal of media and communication research*, 30(56):18, 6 2014. doi: 10.7146/mediekultur.v30i56.8951.

[126] Bruno Latour. From realpolitik to dingpolitik. *Making things public: Atmospheres of democracy*, 2005.

[127] Edward J Lawler. An affect theory of social exchange1. *American Journal of Sociology*, 107(2):321–352, 2001.

[128] Windy Yvonne Lawrence and Benjamin R Bates. Mommy Groups as Sites for Deliberation in Everyday Speech. *Journal of Public Deliberation*, 10(2), 2014.

[129] Azi Lev-On and Bernard Manin. Happy accidents: Deliberation and online exposure to opposing views. *Online deliberation: Design, research and practice*, pages 105–122, 2009.

[130] John M. Levine and Richard L. Moreland. Progress in small group research. *Annual Review of Psychology*, 41(1):585–634, 1990.

[131] John M. Levine and Richard L. Moreland. Group Socialization: Theory and Research. *European Review of Social Psychology*, 5(1):305–336, 1 1994. doi: 10.1080/14792779543000093.

[132] John M. Levine, Hoon-Seok Choi, and Richard L. Moreland. Newcomer innovation in work teams. *Group creativity: Innovation*, 2003.

[133] Peter Levine, Archon Fung, and John Gastil. Future directions for public deliberation. *Journal of Public Deliberation*, 1(1), 2005.

[134] Greg Little, Lydia B Chilton, Max Goldman, and Robert C Miller. Turkit: human computation algorithms on mechanical turk. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 57–66, 2010.

[135] Matthew T. Loveland and Delia Popescu. Democracy on the web. *Information, Communication & Society*, 14(5):684–

703, 8 2011. doi: 10.1080/1369118X.2010.521844. URL http://www.tandfonline.com/doi/abs/10.1080/1369118X.2010.521844.

[136] Anders Sundnes Løvlie, Karoline Andrea Ihlebæk, and Anders Olof Larsson. User Experiences with Editorial Control in Online Newspaper Comment Fields. *Journalism Practice*, pages 1–20, 3 2017. doi: 10.1080/17512786.2017.1293490.

[137] Setha Low and Neil Smith. *The politics of public space*. Routledge, 2013.

[138] Carolyn J Lukensmeyer. Key Challenges Facing the Field of Deliberative Democracy. *Journal of Public Deliberation*, 10(1), 2014.

[139] Carolyn J Lukensmeyer, Joe Goldman, Steven Brigham, J Gastil, and P Levine. A town meeting for the twenty-first century. *The deliberative democracy handbook: Strategies for effective civic engagement in the twenty-first century*, pages 154–163, 2005.

[140] Rousiley C. M. Maia and Thaiane A. S. Rezende. Respect and Disrespect in Deliberation Across the Networked Media Environment: Examining Multiple Paths of Political Talk. *Journal of Computer-Mediated Communication*, 21(2):121–139, 3 2016. ISSN 10836101. doi: 10.1111/jcc4.12155. URL http://doi.wiley.com/10.1111/jcc4.12155.

[141] Meethu Malu, Nikunj Jethi, and Dan Cosley. Encouraging personal storytelling by example. *Proceedings of the 2012 iConference*, 2012. URL http://dl.acm.org/citation.cfm?id=2132309.

[142] Edith Manosevitch, Nili Steinfeld, and Azi Lev-On. Promoting online deliberation quality: cognitive cues matter. *Information, Communication & Society*, 17(10):1177–1195, 11 2014. doi: 10.1080/1369118X.2014.899610.

[143] Jane Mansbridge. Conflict in a New England Town Meeting. *The Massachusetts Review*, 1976.

[144] Jane Mansbridge. Deliberative polling as the gold standard. *The Good Society*, 2010.

[145] Jane Mansbridge, Janette Hartz-Karp, Matthew Amengual, and John Gastil. Norms of deliberation: An inductive study. 2006.

[146] Jane Mansbridge, James Bohman, Simone Chambers, David Estlund, Andreas Føllesdal, Archon Fung, Cristina Lafont, Bernard Manin, and others. The Place of Self-Interest and the Role of Power in Deliberative Democracy. *Journal of political philosophy*, 18(1):64–100, 2010.

[147] David Martin, Benjamin V Hanrahan, Jacki O'Neill, and Neha Gupta. Being a turker. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 224–235, 2014.

[148] Paula McAvoy and Diana Hess. Classroom Deliberation in an Era of Political Polarization. *Curriculum Inquiry*, 43(1):14–47, 1 2013. doi: 10.1111/curi.12000.

[149] Joseph E. McGrath. *Groups: Interaction and performance*, volume 14. Prentice-Hall Englewood Cliffs, NJ, 1984.

[150] Joseph E. McGrath and Holly Arrow. The study of groups: past, present, and future. *Personality and Social*, 2000.

[151] Brian McInnis and Gilly Leshed. Lessons learned: Running user studies with crowd workers. *Interactions*, 23(5), 2016. doi: 10.1145/2968077.

[152] Brian McInnis, Dan Cosley, Chaebong Nam, and Gilly Leshed. Taking a HIT: Designing around Rejection, Mistrust, Risk, and Workers' Experiences in Amazon Mechanical Turk. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2271–2282, 2016.

[153] Brian McInnis, Elizabeth Murnane, Dima Epstein, Dan Cosley, and Gilly Leshed. One and Done: Factors affecting one-time contributors to adhoc online communities. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, volume 27, 2016. ISBN 9781450335928. doi: 10.1145/2818048.2820075.

[154] Brian McInnis, Dan Cosley, Eric Baumer, and Gilly Leshed. Effects of Comment Curation and Opposition on Coherence in Online Policy Discussion. In *Proceedings of the 2018 ACM Conference on Supporting Groupwork*, pages 347–358, Sanibel Island, Florida, 2018. ACM.

[155] Brian McInnis, Gilly Leshed, and Dan Cosley. Crafting Policy Discussion Prompts as a Task for Newcomers. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–23, 11 2018. doi: 10.1145/3274390.

[156] Rony Medaglia and Yang Yang. Online public deliberation in China: evolution of interaction patterns and network homophily in the Tianya discussion forum. *Information, Communication & Society*, 20(5):733–753, 5 2017. doi: 10.1080/1369118X.2016.1203974.

[157] John Stuart Mill. *On liberty (1859)*. na, 1975.

[158] Hamideh Molaei. The prospect of civility in Indonesians' online polarized political discussions. *Asian Journal of Communication*, 24(5):490–504, 9 2014. doi: 10.1080/01292986.2014.917116.

[159] Laurence Monnoyer-Smith and Stéphanie Wojcik. Technology and the quality of public deliberation: a comparison between on and offline participation. *International Journal of Electronic Governance*, 5(1):24–49, 2012.

[160] Alfred Moore. Following from the front: Theorizing deliberative facilitation. *Critical Policy Studies*, 6(2):146–162, 7 2012. doi: 10.1080/19460171.2012.689735.

[161] Richard L Moreland and John M Levine. Socialization in small groups: Temporal changes in individual-group relations. *Advances in experimental social psychology*, 15:137–192, 1982.

[162] Richard L Moreland and John M Levine. Group dynamics over time: Development and socialization in small groups. 1988.

[163] Peter Muhlberger. Virtual Agora project report: Deliberated views regarding school consolidation and educational improvements in Pittsburgh. Technical report, Institute for the Study of Information Technology and Society, Pittsburgh, PA, 2005.

[164] Peter Muhlberger and Lori M Weber. Lessons from the Virtual Agora Project: The effects of agency, identity, information, and deliberation on political knowledge. *Journal of Public Deliberation*, 2(1):1–37, 2006.

[165] Jane Mummery and Debbie Rodan. The role of blogging in public deliberation and democracy. *Discourse, Context & Media*, 2(1):22–39, 2013.

[166] Sean A. Munson and Paul Resnick. Presenting diverse political opinions: how and how much. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1457–1466, 2010.

[167] Sean A. Munson, Stephanie Y. Lee, and Paul Resnick. Encouraging Reading of Diverse Political Viewpoints with a Browser Widget. In *ICWSM*, 2013. URL http://dub.uw.edu/djangosite/media/papers/balancer-icwsm-v4.pdf.

[168] Tom Murray, Beverly Park Woolf, Xiaoxi Xu, Stefanie Shipe, Scott Howard, and Leah Wing. Supporting Social Deliberative Skills in Online Classroom Dialogues: Preliminary Results Using Automated Text Analysis *. *LNCS*, 7315:669–671, 2012.

[169] Tom Murray, Lynn Stephens, Beverly Park Woolf, Leah Wing, Xiaoxi Xu, and Natasha Shrikant. Supporting Social Deliberative Skills Online: The Effects of Reflective Scaffolding Tools. In *International Conference on Online Communities and Social Computing*, pages 313–322, 2013. doi: 10.1007/978-3-642-39371-6_36.

[170] Diana C. Mutz. Is Deliberative Democracy a Falsifiable Theory? *Annual Review of Political Science*, 11(1):521–538, 6 2008. doi: 10.1146/annurev.polisci.11.081306.070308.

[171] Elmie Nekmat and William J. Gonzenbach. Multiple opinion climates in online forums: Role of website source reference and within-forum opinion congruency. *Journalism & Mass Communication Quarterly*, 90(4):736–756, 2013. URL http://journals.sagepub.com/doi/abs/10.1177/1077699013503162.

[172] Elisabeth Noelle-Neumann. The spiral of silence a theory of public opinion. *Journal of Communication*, 24(2):43–51, 1974.

[173] Ray Oldenburg. *The great good place: Cafes, coffee shops, bookstores, bars, hair salons, and other hangouts at the heart of a community*. Da Capo Press, 1999.

[174] David M Olson. Legislatures for Post-Conflict Societies. *Democracy and Deep Rooted Conflict: Options for Negotiators. Stockholm: International Institute for Democracy and Electoral Assistance*, 1998.

[175] Hope Olson. *Mapping beyond Dewey's boundaries: Constructing classificatory space for marginalized knowledge domains*. 1998.

[176] Eli Pariser. *The filter bubble: What the Internet is hiding from you*. Penguin, UK, 2011.

[177] Walter C. Parker. Public Discourses in Schools: Purposes, Problems, Possibilities. *Educational Researcher*, 35(8):11–18, 11 2006. doi: 10.3102/0013189X035008011.

[178] Cynthia Peacock, Joshua M Scacco, and Natalie Jomini Stroud. The deliberative influence of comment section structure. *Journalism: Theory, Practice & Criticism*, page 146488491771179, 7 2017. doi: 10.1177/1464884917711791.

[179] Marta Poblet. Towards a Taxonomy of Crowd-civic Systems. 2017. doi: 10.17605/OSF.IO/4DYR7.

[180] Francesca Polletta and John Lee. Is telling stories good for democracy? Rhetoric in public deliberation after 9/11. *American Sociological Review*, 71 (5):699–721, 2006.

[181] Pablo Porten-Cheé and Christiane Eilders. Spiral of silence online: How online communication affects opinion climate percep-

tion and opinion expression regarding the climate change debate. *Studies in Communication Sciences*, 15(1):143–150, 2015. URL `http://www.sciencedirect.com/science/article/pii/S142448961500021`

[182] Benjamin Powers. The human cost of monitoring the internet. *Rolling Stone*, page 9, 9 2017.

[183] Jennifer Preece and Ben Shneiderman. The reader-to-leader framework: Motivating technology-mediated social participation. *AIS Transactions on Human-Computer*, 2009.

[184] Jenny Preece, Blair Nonnecke, and Dorine Andrews. The top five reasons for lurking: improving community experiences for everyone. *Computers in human behavior*, 20(2):201–223, 2004. URL `http://www.sciencedirect.com/science/article/pii/S074756320300087`

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

[186] Stephen D. Reicher, Russell Spears, and Tom Postmes. A Social Identity Model of Deindividuation Phenomena. *European Review of Social Psychology*, 6(1):161–198, 1 1995. doi: 10.1080/14792779443000049.

[187] Henry M. Robert. *Robert's rules of order newly revised*. 2011.

[188] Ian Rowe. Civility 2.0: a comparative analysis of incivility in online political discussion. *Information, Communication & Society*, 18(2):121–138, 2 2015. ISSN 1369-118X. doi: 10.1080/1369118X.2014.940365. URL `http://www.tandfonline.com/doi/abs/10.1080/1369118X.2014.940365`.

[189] Ian Rowe. Deliberation 2.0: Comparing the Deliberative Quality of Online News User Comments Across Platforms. *Journal of Broadcasting & Electronic Media*, 59(4):539–555, 10 2015. doi: 10.1080/08838151.2015.1093482.

[190] David M Ryfe. Narrative and deliberation in small group forums. *Journal of Applied Communication Research*, 34(1):72–93, 2006.

[191] Niloufar Salehi, Lilly C Irani, Michael S Bernstein, Ali Alkhatib, Eva Ogbe, Kristy Milland, and others. 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*, pages 1621–1630, 2015.

[192] Lynn M. Sanders. Against deliberation. *Political theory*, 25(3):347–376, 1997.

[193] Deborah Schiffrin. Meta-Talk: Organizational and Evaluative Brackets in Discourse. *Sociological Inquiry*, 50(3-4):199–236, 7 1980. ISSN 0038-0245. doi: 10.1111/j.1475-682X.1980.tb00021.x. URL `http://doi.wiley.com/10.1111/j.1475-682X.1980.tb00021.x`.

[194] Michael Schudson. Why conversation is not the soul of democracy. *Critical Studies in Media Communication*, 14(4):297–309, 1997.

[195] Douglas Schuler. Online civic deliberation with e-Liberate. *Online deliberation: Design, research, and practice*, pages 293–302, 2009.

[196] Robert L. Scott and Donald K. Smith. The rhetoric of confrontation. *Quarterly Journal of Speech*, 55(1):1–8, 1969.

[197] Bryan C. Semaan, Scott P. Robertson, Sara Douglas, and Misa Maruyama. Social media supporting political deliberation across multiple public

spheres: towards depolarization. *Proceedings of the 17th*, pages 1409–1421, 2014. URL http://dl.acm.org/citation.cfm?id=2531605.

[198] Bryan C. Semaan, Heather Faucett, Scott P. Robertson, Misa Maruyama, and Sara Douglas. Designing Political Deliberation Environments to Support Interactions in the Public Sphere. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, pages 3167–3176, New York, New York, USA, 2015. ACM Press. ISBN 9781450331456. doi: 10.1145/2702123.2702403.

[199] Pamela J. Shoemaker, James W. Tankard, and Dominic L. Lasorsa. *How to build social science theories*. Sage Publications, 2003.

[200] Buckingham Shum and Simon Buckingham Shum. Cohere: Towards Web 2.0 Argumentation How to cite: Cohere: Towards Web 2.0 Argumentation. pages 97–108, 2008.

[201] Herbert W. Simons. Requirements, problems, and strategies: A theory of persuasion for social movements. *Quarterly Journal of Speech*, 56(1):1–11, 1970.

[202] Paolo Spada and James Raymond Vreeland. Who Moderates the Moderators? The Effect of Non-neutral Moderators in Deliberative Decision Making. *Journal of Public Deliberation*, 9(2), 2013.

[203] Marco R Steenbergen, André Bächtiger, Markus Spörndli, and Jürg Steiner. Measuring Political Deliberation: A Discourse Quality Index. *Comparative European Politics*, 1(1):21–48, 3 2003. doi: 10.1057/palgrave.cep.6110002.

[204] Kim Strandberg and Janne Berg. Online Newspapers' Readers' Comments - Democratic Conversation Platforms or Virtual Soapboxes? *Comunicação e Sociedade*, 23(1):132, 2013. doi: 10.17231/comsoc.23(2013).1618.

[205] Kim Strandberg and Janne Berg. Impact of Temporality and Identifiability in Online Deliberations on Discussion Quality: An Experimental Study. *Javnost - The Public*, 22(2):164–180, 2015. doi: 10.1080/13183222.2015.1041230.

[206] Marcy Strauss. Sequestration. *American Journal of Criminal Law*, 24, 1996.

[207] Jennifer Stromer-Galley. Measuring deliberation's content: A coding scheme. *Journal of public deliberation*, 3(1), 2007.

[208] Jennifer Stromer-Galley and Anna M. Martinson. Coherence in political computer-mediated communication: analyzing topic relevance and drift in chat. *Discourse & Communication*, 3(2):195–216, 5 2009. doi: 10.1177/1750481309102452.

[209] Jennifer Stromer-Galley and Peter Muhlberger. Agreement and Disagreement in Group Deliberation: Effects on Deliberation Satisfaction, Future Engagement, and Decision Legitimacy. *Political Communication*, 26(2):173–192, 2009. doi: 10.1080/10584600902850775.

[210] Jennifer Stromer-Galley and Alexis Wichowski. Political discussion online. In *The handbook of internet studies*, chapter 11, page 168. 2011.

[211] Jennifer Stromer-Galley, Lauren Bryant, and Bruce Bimber. Context and Medium Matter: Expressing Disagreements Online and Face-to-Face in Political Deliberations. *Journal of Public Deliberation*, 11(1), 2015.

[212] Natalie Jomini Stroud, Joshua M. Scacco, Ashley Muddiman, and Alexander L. Curry. Changing Deliberative Norms on News Organizations' Facebook Sites. *Journal of Computer-Mediated Communication*, 20(2):188–203, 3 2015. doi: 10.1111/jcc4.12104.

[213] Colleen Stuart, Laura Dabbish, Sara Kiesler, Peter Kinnaird, and Ruogu Kang. Social transparency in networked information exchange: a theoretical framework. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 451–460, 2012.

[214] Abhay Sukumaran, Stephanie Vezich, Melanie McHugh, and Clifford Nass. Normative influences on thoughtful online participation. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, page 3401, New York, New York, USA, 2011. ACM Press. ISBN 9781450302289. doi: 10.1145/1978942.1979450. URL `http://dl.acm.org/citation.cfm?doid=1978942.1979450`.

[215] Cass R Sunstein. The law of group polarization. *Journal of political philosophy*, 10(2):175–195, 2002.

[216] Cass R Sunstein. Deliberating groups versus prediction markets (or Hayek's challenge to Habermas). *Episteme*, 3(03):192–213, 2006.

[217] Sharon E. Sutton and Susan P. Kemp. Children as partners in neighborhood placemaking: lessons from intergenerational design charrettes. *Journal of Environmental Psychology*, 22(1):171–189, 2002.

[218] Dennis F. Thompson. Deliberative Democratic Theory and Empirical Political Science. *Annual Review of Political Science*, 11(1):497–520, 6 2008. doi: 10.1146/annurev.polisci.11.081306.070555.

[219] W. Ben Towne and James D. Herbsleb. Design considerations for online deliberation systems. *Journal of Information Technology & Politics*, 9(1):97–115, 2012.

[220] Richard van der Wurff, Knut De Swert, and Sophie Lecheler. News Quality and Public Opinion: The Impact of Deliberative Quality of News Media on Citizens' Argument Repertoire. *International Journal of Public Opinion Research*, 8 2016. doi: 10.1093/ijpor/edw024.

[221] Vasilis Verroios and Michael S. Bernstein. Context trees: Crowdsourcing global understanding from local views. In *Second AAAI Conference on Human Computation and Crowdsourcing*, 2014.

[222] Jing Wang, Clement T. Yu, Philip S. Yu, Bing Liu, and Weiyi Meng. Diversionary comments under political blog posts. In *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12*, page 1789, New York, New York, USA, 2012. ACM Press. ISBN 9781450311564. doi: 10.1145/2396761.2398518. URL `http://dl.acm.org/citation.cfm?doid=2396761.2398518`.

[223] Etienne Wenger and Jean Lave. Legitimate peripheral participation in communities of practice. *Supporting lifelong learning*, 2002.

[224] Anita W. Woolley, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone. Evidence for a collective intelligence factor in the performance of human groups. *Science (New York, N.Y.)*, 330(6004):686–8, 10 2010. doi: 10.1126/science.1193147. URL `http://www.ncbi.nlm.nih.gov/pubmed/20929725`.

[225] Scott Wright and John Street. Democracy, deliberation and design: the case of online discussion forums. *New media & society*, 9(5):849–869, 2007.

[226] Scott Wright, Todd Graham, and Daniel Jackson. Third Space , Social Media and Everyday Political Talk.

[227] Lu Xiao and Nicole Askin. What influences online deliberation? A wikipedia study. *Journal of the Association for Information Science and Technology*, 65(5):898–910, 5 2014. doi: 10.1002/asi.23004.

[228] Wenjie Yan, Gayathri Sivakumar, and Michael A. Xenos. It's not cricket: Examining political discussion in nonpolitical online space. *Information, Communication & Society*, pages 1–17, 6 2017. doi: 10.1080/1369118X.2017.1340499.

[229] Thomas Zerback and Nayla Fawzi. Can online exemplars trigger a spiral of silence? Examining the effects of exemplar opinions on perceptions of public opinion and speaking out. *New Media & Society*, 2016. URL `http://nms.sagepub.com/content/early/2016/01/22/1461444815625942.`

[230] Amy X. Zhang, Lea Verou, and David Karger. Wikum: Bridging Discussion Forums and Wikis Using Recursive Summarization. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 2082–2096. ACM, 2017.

[231] Marc Ziegele, Timo Breiner, and Oliver Quiring. What Creates Interactivity in Online News Discussions? An Exploratory Analysis of Discussion Factors in User Comments on News Items. *Journal of Communication*, 64 (6):1111–1138, 12 2014. doi: 10.1111/jcom.12123.

[232] Marc Ziegele, Mathias Weber, Oliver Quiring, and Timo Breiner. The dynamics of online news discussions: effects of news articles and reader comments on users' involvement, willingness to participate, and the civility of their contributions. *Information, Communication & Society*, pages 1–17, 5 2017. doi: 10.1080/1369118X.2017.1324505.

[233] Jonathan Zittrain. Work the new digital sweatshops. *Newsweek, December*, 8, 2009.