TOWARDS COMPUTATIONAL METHODS FOR PROACTIVELY SUPPORTING HEALTHIER ONLINE DISCUSSIONS

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

Jonathan Pei-Wah Chang

May 2024

TOWARDS COMPUTATIONAL METHODS FOR PROACTIVELY SUPPORTING HEALTHIER ONLINE DISCUSSIONS

Jonathan Pei-Wah Chang, Ph.D.

Cornell University 2024

One of the biggest problems facing online platforms today is the prevalence of so-called "toxic" behavior, such as personal attacks, harassment, and general incivility. While a common computational approach for addressing this problem has been developing algorithms to detect toxicity, we argue that this approach reflects an overly narrow view of online community governance, catering specifically to the use case of platform-driven, centralized content moderation, while overlooking an equally important perspective: that of the *communities* of ordinary users who interact on these platforms. Therefore, this dissertation takes on the following question: how can technology support members of online communities in having healthier interactions, and thereby proactively prevent toxicity from taking root?

We take a combined social and technical approach to answering this question. From the social perspective, we begin with a close examination of existing practices of online community governance: drawing from literature in diverse fields ranging from computer science to sociology, law, and political science, we identify concrete ways in which online communities proactively prevent toxicity and promote pro-social norms, and conduct interviews to gain more qualitative insights. These insights guide our technical approach: inspired by interview participants' explanations of how they can intuitively tell whether a conversation might later derail into toxicity, we formalize such *derailment fore-*

casting as a novel computational task and argue that solving it requires a new class of *conversational forecasting models*. Finally, bringing together the technical and social aspects, we develop a first-of-its-kind concrete implementation of a conversational forecasting model and evaluate it via an "in-the-wild" user study involving ordinary users in a real online community.

We conclude by looking back on our findings thus far and comparing them with our higher-level, long-term goals for this work. From this comparison, we identify current shortcomings and unanswered questions that should be tackled in future work, and pull in insights from recent developments in machine learning, natural language processing, and computational social science to build a concrete roadmap of next steps.

BIOGRAPHICAL SKETCH

Jonathan P. Chang was born and raised in San Jose, California, and prior to starting his Ph.D. he actually spent his entire life in California, having attended Harvey Mudd College for his B.S. in Computer Science. Moving to Ithaca for his Ph.D. therefore represented the first time he ever lived in a location with real seasons and snow. Further research is needed on how much this shaped his subsequent research agenda and worldview—though it is worth noting that his decision to return to Harvey Mudd College for a faculty position may constitute a vote of confidence in the California lifestyle.

This document is dedicated to all my fellow terminally-online netizens who have ever unintentionally found themselves in the midst of an online conversation gone awry.

ACKNOWLEDGEMENTS

The only reasonable way to begin this long list of acknowledgements is by recognizing my advisor, Cristian Danescu-Niculescu-Mizil, and the key role he played in my formative years as a researcher. I have long been open about the fact that, as a first-year Ph.D. student coming straight from undergrad, I had absolutely no idea what I wanted to do. My prior research experience was limited, I had no concrete research vision beyond some vague notions of "technology plus society", and I was almost completely disconnected from Computer Science academia—a fact that came into stark relief in a particularly embarrassing moment at Visit Days where I failed to recognize the name Jon Kleinberg. Yet after a chance meeting over lunch on the first day of my first semester, Cristian evidently saw some potential in me, and invited me to his seminar on NLP and computational social science—an entirely coincidental arrangement that ended up forming the core of my eventual research agenda.

Speaking of that seminar, I would be remiss in not mentioning the many other amazing people, students and faculty alike, that I met there. There's Lillian Lee, whose ability to pull keen and novel insights about a paper out of seemingly thin air I still aspire to emulate; Justine Zhang, who taught me everything I know about how to keep research code well-structured and how to give effective research talks; Tom Davidson, who served as my first exposure to the world of content moderation scholarship outside of computer science; Xanda Schofield, whose trajectory as a fellow Mudd alum who ended up doing computational social science research on the path to a teaching career is exactly what I strove to follow; and Ana Smith, with whom I've had an uncountable number of fun conversations (and debates) about varied topics both related and unrelated to research.

Of course, there are additional avenues through which I've met peers, mentors, and collaborators here at Cornell. Throughout what is now the Cornell Bowers College of Computing and Information Sciences, I've met a number of helpful faculty who have either directly collaborated with me or otherwise provided guidance and mentoring, including Tom Ristenpart, Karen Levy, David Mimno, Matt Wilkens, and Adrian Sampson. I would also like to extend my sincerest appreciation to the many staff members who help the department and College run smoothly day to day, including Becky Stewart for her frankly superhuman helpfulness with navigating the complex procedures of the Ph.D. program, Janeen Orr and Josh Hunt for their diligent efforts in supporting those of us who work on the second floor of Gates Hall, and Chris Fouracre from IT for maintaining our research servers. And then, of course, there are my fellow Ph.D. students; I'd particularly like to recognize my many lab- and office-mates over the years, including Liye Fu, Maria Antoniak, Laure Thompson, Jack Hessel, Greg Yauney, Roz Thalken, Rebecca Hicke, Anna Choi, and Andrea Wang. I'd also like to acknowledge the emotional support of many friends from outside my immediate research circle, including Claire Liang, Cheng Perng Phoo, Dietrich Geisler, and Sam Havron.

Moving beyond the small circle of Gates Hall, I also want to recognize the role of two Cornell institutions besides Bowers CIS that have shaped my experiences here. In the Department of Philosophy, I'd like to thank Julia Markovits for agreeing to act as minor advisor for some random computer science student she'd never met, and W. Starr for some insightful and inspiring conversations that ended up being more directly relevant to my research than I may have originally anticipated. I also thank the Cornell Center for Social Sciences (CCSS) for supporting me for two years as part of their Data Science Fellowship. I es-

pecially thank the current and former members of CCSS staff who were closely involved in that fellowship, including Peter Enns, Drew Margolin, Katie Anderson, Claudia Von Vacano, Lynda Kellam, Lynn Martin, Florio Arguillas, Jacob Grippin, and Cassian D'Cunha. And of course there is my "second cohort": the other amazing CCSS fellows I've worked with, including Kimberly Williamson, Yolanda Xue, Aspen Omapang, Remy Stewart, and Aishat Sadiq.

Support for my journey also came from areas outside Cornell. I would like to thank Wikimedia Foundation Research for supporting some of my earliest work, and for inviting me to present at two WMF research showcases—early experiences that helped me build up my confidence in giving research talks. I particularly thank my main contacts at WMF research, Dario Taraborelli and Lelia Zia. I also thank the Facebook (now Meta) Core Data Science team for offering me the opportunity to intern with them and get a taste of industry research. I am grateful to my internship mentor Justin Cheng, and to Lada Adamic, Israel Nir, Alex Dow, Ashish Gupta, Karen Jusko, Alex Leavitt, and Moira Burke for their advice and feedback throughout.

In addition to the previously mentioned support from WMF Research and the CCSS Data Science Fellowship, the research described in this dissertation was also supported in part by NSF CAREER Award IIS-1750615, NSF Grants SES-1741441 and IIS-1910147, a Google Faculty Research Award, a Crowd-Flower AI for Everyone Award, and a Google Cloud Platform research credit.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

AN INTRODUCTION TO ONLINE COMMUNITY GOVERNANCE AND MODERATION

1.1 Online Toxicity: The Problem and Existing Solutions

> "Two redditors enter, two redditors leave after accomplishing nothing but getting angry on the internet."[1]
>
> – Anonymous Reddit user

The above quote from a Reddit comment thread illustrates what has long since become conventional wisdom: social media and the internet, for all they have done to connect us, are at the same time filled with toxicity. With growing awareness of this problem has also come increasing pressure on major online platforms—coming from all corners of society, from politicians to the media to community groups—to do something to stem the spread of toxic content. If we were to ask the average person to imagine the standard process of dealing with a toxic post on Reddit (or Facebook, Twitter, etc.), the scenario in their head would probably look something like this:

> A moderator becomes aware of toxic content, and then they take it down.

Indeed, the centralized moderation process described in the above scenario is the most well-known approach to online community governance, as has been

[1] Comment source: `https://www.reddit.com/r/SubredditDrama/comments/19fbb54/comment/kjiihfc/`

covered in great detail by—among others—Gillespie (2018)'s seminal work on content moderation. But is this whack-a-mole style approach to toxic content, wherein moderators must constantly react to a never-ending stream of toxic content, truly the end-all-be-all of online community governance? Can we imagine alternative scenarios?

Let us begin by proposing one way we can break down the above scenario into its fundamental components. First, there is the question of **who** is involved: beyond the now-stereotypical concept of a designated content moderator (who is, in the popular conception, almost always a professional worker employed or contracted by the platform owner), are there other figures who might play a key role in combating toxicity? Second, there is the question of **what** actions the involved figures (be they moderators or somebody else) take in order to combat toxicity. Finally, there is the question of **when** the above actions take place.

If we apply this framework to the stereotypical moderation scenario we sketched above, this is what we get:

> [A moderator **(WHO)**] becomes aware of toxic content, [and then **(WHEN)**] they [take it down **(WHAT)**]

In other words, this scenario arises from assigning "professional moderators" as the **who**, "content removal/takedown" as the **what**, and "sometime after the toxic content was posted" as the **when**. But a survey of the vast literature on online community governance reveals a much larger landscape of other combinations of **who/what/when** yielding a variety of alternative approaches to managing toxicity in online communities—some of which are already practiced, and others of which are more hypothetical yet ripe with potential.

1.1.1 The "Who" Dimension: Actors Involved in Community Governance

Today, many online platforms take a *platform-driven* approach to governance, where platform operators directly employ or contract workers to review potentially objectionable content and remove it if needed (Gillespie, 2018). This is arguably the model of moderation that the lay audience is most familiar with, as it has been adopted by the most prominent platforms such as Facebook and Twitter, and has been a driving force behind high-profile moderation cases such as Reddit's 2015 mass ban of hate communities (Chandrasekharan et al., 2017). However, this strategy also suffers from key weaknesses. Chief among these is the problem of *scale*: the large amount of content being generated on major online platforms makes it infeasible for moderators to handle all content needing review in a timely manner (Gillespie, 2020), and results in a high workload and stress for the moderators (Roberts, 2014).

Though the platform-driven approach may be dominant in today's Web, its ascendancy was by no means a foregone conclusion: early online communities, with their decentralized ethos, tended to instead prefer a bottom-up, *community-driven* model (Dibbell, 2005; Lampe and Resnick, 2004). As of late, community-driven governance has seen a renewed surge in interest in light of the shortcomings of platform-driven moderation (Brewer et al., 2020; Seering, 2020; Zhang et al., 2020a), and it remains the method of choice in smaller, interest-specific communities—for example, Twitch livestream communities (Lo, 2018; Cai and Wohn, 2019) and the topical groups on Reddit known as "subreddits" (Dosono and Semaan, 2019; Chandrasekharan et al., 2018; Gilbert, 2020). Community-driven governance practices can be further subdivided as roughly falling into

two categories: volunteer moderators and ordinary user involvement.

One common approach to community-driven governance mimics the centralized model of platform-driven moderation, granting the authority to review and remove content to a core group of *volunteer moderators*, who are not platform employees but rather regular community members who have stepped up to the task (Dosono and Semaan, 2019; Geiger and Ribes, 2010; Lo, 2018; Seering, 2020; Wohn, 2019). While volunteer moderators are conceptually similar to platform-employed moderators in terms of their administrative powers and workflow, their status as actual members of the communities they moderate can be a unique advantage: they may receive a higher level of trust and connection from the community, unlike platform-employed moderators who are seen as outsiders (Seering, 2020), and their inside knowledge of community norms and dynamics can help them negotiate harder, more nuanced disputes (Turnbull, 2018; Chandrasekharan et al., 2018; Shahid et al., 2024). On the other hand, like their platform-employed counterparts, volunteer moderators face the problem of scale, and the resulting problems of overwork and stress are exacerbated by the fact that volunteer moderators are doing this work in their free time, not as their full-time job (Dosono and Semaan, 2019; Wohn, 2019).

As such, online communities have sought strategies to mitigate the problem of uncivil behavior outside the framework of centralized moderation, thereby decreasing the burden on moderators. This has led to a second family of approaches which aim to involve *ordinary users* in the everyday governance of their communities (Kiesler et al., 2012; Seering, 2020). In contrast to moderator-centric approaches, tools and policies that involve ordinary users in the governance process tend to be less authoritative in order to prevent the risk of misuse

(e.g., ordinary users should not have the ability to remove someone else's content), and are instead softer and smaller in scope. One particularly common way to involve ordinary users in community governance is to allow them to *vote* on whether a piece of content constitutes a valuable contribution to the community; content that receives too many negative votes can then be automatically de-prioritized or hidden (Lampe and Resnick, 2004; Mamykina et al., 2011; Chandrasekharan et al., 2018; Papakyriakopoulos et al., 2023). Even more limited in scope is the personalized *blocklist* (Geiger, 2016; Jhaver et al., 2018b), which allows users to specify that they do not want to see content from specific other users in their personal feed, but does not otherwise impact that content on the rest of the platform. Finally, as a bridge between ordinary users and moderators, some platforms allow users to *flag* content that they find objectionable; this action does not have any immediate effect on its own but places the flagged content in a queue for moderators to make a final decision on it (Crawford and Gillespie, 2016; Kou and Gui, 2021).

1.1.2 The "When" Dimension: How Soon Can We Take Action?

While content removal is a widely studied and practiced strategy, it also comes with an inherent weakness: because it involves taking action against toxic content that has *already* been posted, the offending content has an opportunity to be seen and to spread before moderators are able to take action (if they ever do at all). In the meantime, this can harm users exposed to the toxic content, in addition to harming the platform by preventing or distracting from productive discussions. Although this common *reactive* paradigm is admittedly better than doing nothing at all, some have advocated that a more effective way to protect

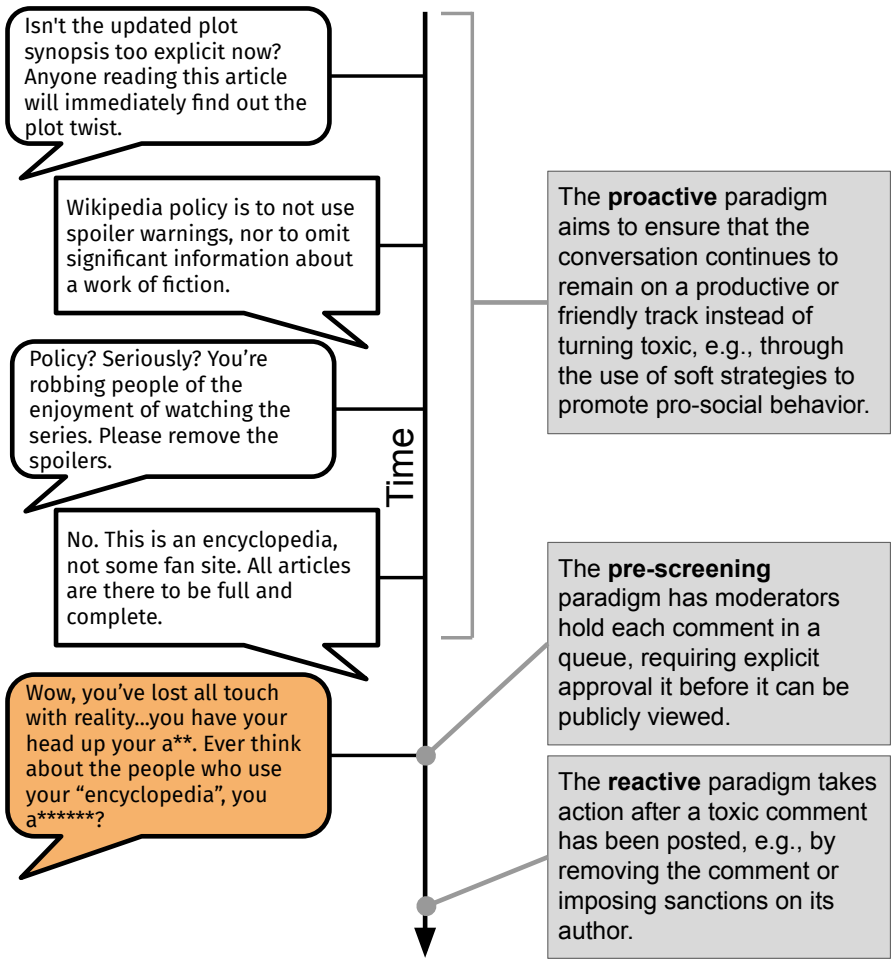


Figure 1.1: Three paradigms of online community governance that vary along the **when** dimension, exemplified in the context of a conversation between two Wikipedia editors that eventually derails into a personal attack (orange).

online communities from harm is to reduce the amount of toxic content that gets posted in the first place (Kiesler et al., 2012; Grimmelmann, 2015). This goal has motivated work on a number of alternative paradigms which involve taking action earlier in the lifecycle of an online interaction (illustrated in Figure 1.1).

In the current media environment, one alternative to the reactive paradigm which has occasionally been attempted is *pre-screening*, which stipulates that content must be reviewed and explicitly approved by moderators before it appears on the platform (Kiesler et al., 2012). This approach, hearkening back

to the days of traditional pen-and-paper media, is still employed by a handful of platforms such as the *New York Times* comment section.[2] Additionally, there has been work on hybrid automatic/human systems for comment pre-screening (Park et al., 2016). However, most platforms avoid this strategy because it raises a host of practical issues. Pre-screening is highly labor intensive, and scales poorly as a platform grows: for example, even with the help of algorithmic pre-screening, the *New York Times* currently only allows comments on top stories for 8 hours during weekdays. Moreover, pre-screening prevents real-time interaction between users on a platform by introducing a delay between users submitting content and that content appearing on the platform while moderators review it. Finally, pre-screening has been subjected to criticism on the grounds of suppressing free speech (Gillespie, 2018).

Consequently, recent work in online community governance has advocated for going even further: online communities should aim to discourage the creation of toxic content in the first place, rather than waiting to eventually remove it or screen it out. This goal can colloquially be thought of as helping users to bear in mind the human being on the other side of the screen, by fostering community norms that reproduce or stand in for the norms and social signals that typically serve to discourage toxicity in face-to-face settings (Bicchieri, 2016) (as famously depicted in humorous fashion by the comic in Figure 1.2). This paradigm is sometimes described as *proactive* (Lo, 2018; Seering, 2020; Seering et al., 2017; Cai et al., 2021), in contrast to the previously described reactive paradigm which covers strategies like post-hoc removal of toxic content.

In practice, perhaps the most common type of proactive strategy currently employed is the use of deliberate choices in platform design aimed at promoting

[2] `https://help.nytimes.com/hc/en-us/articles/115014792387-Comments`

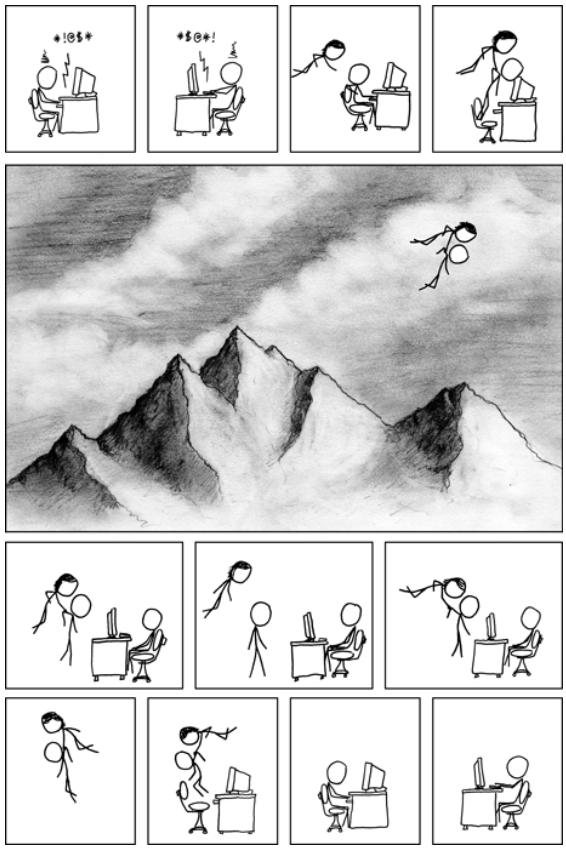


Figure 1.2: The classic comic "Internet Argument" from cartoonist Randall Munroe's long-running *xkcd* humorously illustrates the big-picture goal of proactive approaches to online community governance: reproducing, in the online setting, the social norms that make offline toxicity much rarer. (Source: `https://xkcd.com/438/`, licensed under CC BY-NC 2.5)

pro-social behaviors. These have a long and established history in social computing; now-common design choices such as activity indicators (Erickson and Kellogg, 2000) and explicitly listed rules (Kiesler et al., 2012) were initially developed as measures to encourage the development and adoption of pro-social norms within online communities. More recent developments in this direction include limitations on community size or rate of participation (Grimmelmann, 2015), codes of conduct designed with community input (Li et al., 2021), and user interfaces that draw on insights from psychology to prime users towards empathy (Taylor et al., 2019).

That said, static design choices can only go so far, and so as platforms have grown and evolved, they have developed more dynamic strategies for proactive community management. For instance, a natural development from static listing of rules involves sending explicit reminders of community rules specifically in high-impact situations, such as when welcoming newcomers (Halfaker et al., 2011b; Seering et al., 2019b). As a further step from this, recent work has looked at how volunteer moderators can model good behavior in their own interactions, as a way of implicitly signaling to the community what proper behavior looks like (Jagannath et al., 2020; Seering et al., 2017).

1.1.3 The "What" Dimension: What Action Can be Taken?

The standard combination of "professional moderators" in the **who** dimension and "after toxic content is posted" in the **when** dimension somewhat constrains the third dimension, that is the action space of **what** can be done. While the most common—or at least the most prominent—answer is content removal, other reactive strategies often employed by professional moderators include limiting the visibility of toxic content rather than removing it outright (Lampe and Resnick, 2004), and temporarily or permanently banning the authors of toxic content from the platform (Chang and Danescu-Niculescu-Mizil, 2019a; Jhaver et al., 2021). Some platforms also combine these two approaches and limit content visibility on a user-specific basis, allowing the blocking of content from known bad actors (Jhaver et al., 2018b)—a practice sometimes referred to as "shadowbanning" (Delmonaco et al., 2024).

Volunteer moderators, on the other hand, arguably have a larger range of ac-

tions available to them, including a number of proactive strategies. This arises from the fact that—unlike most professional moderators—volunteer moderators are by definition part of their communities, and many will continue to participate in conversations and other informal interactions (Wohn, 2019; Lo, 2018). Volunteer moderators must therefore balance their "dual identities" as both regular community members and authority figures. Different ways of managing this balance will result in different conceptions of one's role and purpose as a moderator; in interviews, volunteer moderators have described their work with metaphors that range from the formal ("police", "governor", "manager") to the informal ("team member", "facilitator", "adult in the room") (Seering et al., 2020). This diversity in attitudes towards moderation naturally leads to a diversity in employed methodology. While many volunteer moderators can and do wield the authority to take harsh reactive measures such as removing content (Gilbert, 2020; Jhaver et al., 2019a) or suspending users (Lo, 2018; Chang and Danescu-Niculescu-Mizil, 2019a), they often express a preference for softer, social approaches to *proactively* keep the community in line (Seering et al., 2019b). Examples of such soft strategies include publicly modeling good behavior (Seering et al., 2017; Cai et al., 2021), educating users about the rules (Cai and Wohn, 2019), and mediating disputes (Billings and Watts, 2010).

Meanwhile, ordinary users by definition lack access to formal moderation tools (other than specifically user-facing tools like reporting mechanisms and blocklists) and therefore can only rely on soft strategies if they want to combat or prevent toxicity. Yet despite this limitation, the ordinary users within an online discussion are arguably the most well-positioned to prevent toxicity within the discussion, as they are the ones steering the direction of the discussion, and the fact that they are already participating may allow them to act more quickly

than a moderator coming in from the outside. One notable way in which ordinary users in online discussions have leveraged their unique positioning to fight toxicity is "counterspeech", in which users combat hateful posts and comments by posting a reply or competing post that aims to rebut the hateful narrative (Chung et al., 2019; Mathew et al., 2019) and proactively prevent other users from turning hateful (He et al., 2022). Recognizing the privileged role ordinary users may play in keeping their discussions on a healthy track, an emerging line of research is looking into how to design technical and interface-level *interventions* to encourage users to keep their discussions civil (Kriplean et al., 2012b; Seering et al., 2019a; Argyle et al., 2023).

1.2 Our Contributions

Today's approaches to applying technology to help with online community governance largely take the form of algorithms for detecting toxic content (as we will discuss in detail below in Section 1.2.1), which can be used to supplement professional moderators in taking down such content reactively. Yet as we have seen in Section 1.1, there is far more to the landscape of online community governance than just reactive strategies undertaken by professional moderators. It is worth asking, then, whether algorithmic tools have the potential to empower other types of paradigms—in particular, proactive strategies that can be undertaken by both volunteer moderators and ordinary users.

This dissertation, then, is focused on addressing this question. We broadly proceed in three steps (visualized in Figure 1.3): first, we must more thoroughly examine the workflow of both volunteer moderators and ordinary users who

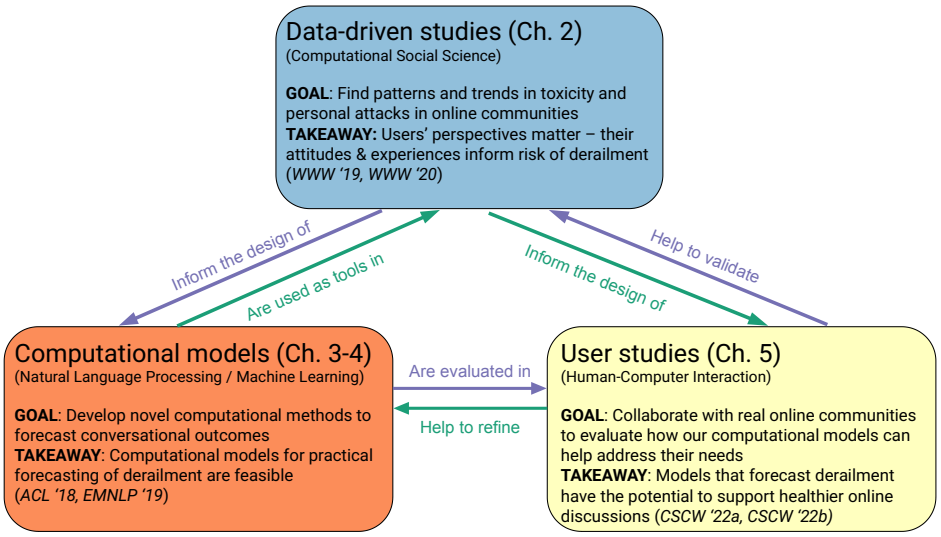


Figure 1.3: A visual overview of our joint social and technical research agenda, showing the different components drawing from different subfields, and how these components interact with each other.

engage in various proactive strategies for preventing toxicity, so that we can identify concrete needs that algorithmic tools could help with. Second, we must determine what algorithmic methods can be deployed to address those needs—and where technological gaps exist, we must fill them in by innovating new methods. Finally, we must evaluate the impact of the resulting algorithmic solutions on actual online communities, with a focus on understanding to what extent they actually meet the needs of volunteer moderators and ordinary users, and to what extent they still fall short.

1.2.1 Motivation: The Promise and Limitations of Algorithmic Assistance

Multiple studies on content moderation have identified a problem of scale: even if toxicity is a small fraction of all content that gets posted, the sheer size of modern online platforms, together with the relatively small number of moderators present on most platforms, makes it infeasible for human moderators to keep up with all the toxic content that gets posted (Dosono and Semaan, 2019; Lo, 2018; Wulczyn et al., 2017; Gillespie, 2020). This has led to mental strain and burnout among moderators (Dosono and Semaan, 2019) and has directly inspired calls for the development of technological assistance to reduce the burden on human moderators (Wohn, 2019). As Gillespie (2020) writes, "the strongest argument for the automation of content moderation may be that, given the human costs, there is simply no other ethical way to do it, even if it is done poorly". Technological responses to this call have ranged in complexity: basic tools include simple word-based filters (Wohn, 2019; Lo, 2018; Chancellor et al., 2016) and blocklists (Jhaver et al., 2018b), while more advanced systems attempt to use machine learning and natural language processing techniques to automatically identify toxic content (Wulczyn et al., 2017; Nobata et al., 2016; Gambäck and Sikdar, 2017).

Regardless of the choice of technical backend, most algorithmic tools are specifically targeted towards (professional or volunteer) moderators within reactive or pre-screening paradigms. A common use case is to apply the filter or classifier to content that has already been submitted for public posting; while in rare cases this can be applied as a pre-screening approach where the filter automatically blocks certain submitted content from getting posted (usually in-

volving high-precision filters that look for hand-chosen terms known to cause problems in a specific micro-community) (Wohn, 2019; Lo, 2018), the more common application is to use the filter or classifier to flag content for review by human moderators (allowing the content to stay public in the meantime) (Lo, 2018; Chandrasekharan et al., 2019; Jhaver et al., 2019b; Geiger and Ribes, 2010), who may then decide to (reactively) remove the content and/or impose other penalties.

Although technology companies have been publicly optimistic about the potential for such algorithmic tools to "fix" toxicity and other problems facing moderators (Katzenbach, 2021), there is plenty of cause for doubt that today's algorithmic tools are up to the task. At a basic technical level, audits of these tools have shown they still have much room for improvement when it comes to accurately detecting toxic content (Cao et al., 2023). To make matters worse, such inaccuracy is not equitably distributed, with multiple experiments having demonstrated that algorithmic toxicity classifiers reproduce and amplify societal biases against underrepresented and underprivileged groups (Davidson et al., 2017; Park et al., 2018; Wiegand et al., 2019; Sap et al., 2019; Zhang et al., 2020b). But this is more than just a technical issue that can be solved with better models or more data—the very task of "toxicity detection" faces problems inherent to its setup. When thinking about the higher-level goal of creating healthier online communities, it becomes clear that toxicity is only one part of the problem; more subtle behaviors such as dogwhistles (Mendelsohn et al., 2023), faux-good-faith questioning a.k.a. "sealioning" (Johnson, 2017), and other kinds of "veiled" attacks (Han and Tsvetkov, 2020) all threaten community health and create a less welcoming and inclusive environment, despite not being overtly toxic. Indeed, "toxicity" is not even a well-defined notion, be-

ing heavily dependent on cultural and social context (Sheth et al., 2022), making it unclear whether it is even possible to rigorously formulate toxicity detection as a straightforward labeling task.

What, then, is to be done? We believe that the answer lies in going beyond the perspective of reactive, platform-driven moderation. As we have outlined, online community governance encompasses a much wider world of practices, and it stands to reason that if our goal is to improve online communities, then such work should naturally start from a community-level perspective (Hasinoff and Schneider, 2022). To be clear, we do not claim this to be a cure-all for the problems described above—algorithmic systems will still suffer from problems like bias and misalignment with societal values, regardless of whether they are deployed in a reactive platform-driven context or a proactive community-driven one. But when we consider what these different approaches offer in terms of responding to and handling cases of error and bias, the community-driven approach appears to offer clear benefits: it distributes power and responsibility across the whole community rather than concentrating it in the hands of platform owners, thereby having the potential to accommodate more nuance and to rapidly adapt to shifting social norms and values. Through our joint technical and social approach (Figure 1.3), this dissertation aims to lay the early groundwork for such an approach, identifying the problems that need to be solved and innovating new technical approaches to solving them.

1.2.2 Organization

The rest of this dissertation is organized along the aforementioned three steps.

In **Chapter 2**, we take a deeper look at proactive strategies for preventing toxic outcomes, from two perspectives: that of volunteer moderators and that of ordinary users. From the moderators' perspective, we aim to both qualitatively understand how they reason about toxic outcomes and effective strategies to prevent them, and to more quantitatively examine the practical challenges they face in doing so. Likewise, from the users' perspective, we wish to characterize the process by which toxic outcomes arise even among well-intentioned users who legitimately desired to have a friendly or constructive discussion—a process that we refer to as *conversational derailment* and that represents a core focus of our work. We synthesize these findings to arrive at a preliminary picture of what algorithmic assistance would need to be capable of in order to help with proactive strategies—namely, we identify a need for novel algorithmic models that can follow online discussions in real time and track their trajectory, so that discussions at risk of derailing into toxicity can be identified.

In **Chapter 3**, we formalize the modeling task that we had derived from the findings in the preceding chapters. Inspired by foundational work in sociolinguistics, we hypothesize that specific linguistic phenomena may underpin the process of derailment into toxicity, and posit that algorithmically identifying these phenomena may enable successful prediction of future derailment—a novel task within the broader category of "conversational forecasting". Using classical techniques from natural language processing, we establish the feasibility of this task. We then follow up in **Chapter 4** by describing the practical barriers that must be overcome to go beyond the proof of concept and build conversational forecasting models that are actually useful to moderators and ordinary users. We introduce a first-of-its-kind model, CRAFT, that addresses these practical challenges.

Finally, in **Chapter 5**, we describe preliminary work on understanding the role that models like CRAFT can play in a real online community setting. To achieve this, we build user-facing tools powered by CRAFT and evaluate them via user studies, which involved real moderators and users from two online platforms, Wikipedia and Reddit. We show early evidence that this approach has the potential to reduce toxicity and promote pro-social outcomes—while also acknowledging that there remain numerous open questions for future work to tackle, as we lay out in **Chapter 6**.

CHAPTER 2

ONLINE COMMUNITIES' STRATEGIES FOR PROACTIVELY PREVENTING TOXICITY

2.1 Introduction

Addressing toxicity is a major focus for many online communities (Chandrasekharan et al., 2017), as toxicity hinders the exchange of ideas (Arazy et al., 2013) and takes a significant emotional toll on community members who are exposed to it (Ashktorab and Vitak, 2016; Jhaver et al., 2018a). Traditionally, platforms attempt to address this problem through reactive moderation (see 1.1.2), in which volunteers from within the community (Seering, 2020) or professionals employed by the platform operator (Gillespie, 2018) aim to identify and remove "bad actors" and "objectionable content". Substantial efforts are focusing on scaling up this paradigm through automation or algorithmic assistance, an enterprise which has proven to be both technically and ethically challenging (Grimmelmann, 2015; Gillespie, 2020; Katzenbach, 2021; Gorwa et al., 2020).

This common paradigm, however, does not account for the fact that toxic behavior in online communities is not solely the product of "bad actors"—who are generally a minority within their communities (Kumar et al., 2018)—but can instead often emerge from ordinary users when they find themselves in particularly heated or tense situations (Cheng et al., 2017). In fact, in many settings the vast majority of individuals on a platform are *well-intentioned*, in the sense that their purpose for being on the platform is simply to consume interesting content and engage in good faith with other community members (Srinivasan et al., 2019; Gilbert, 2020; Weld et al., 2022). This viewpoint forms

the foundation of the alternative proactive approaches discussed in 1.1.2: given that well-intentioned users do not explicitly intend to engage in toxic behavior, they may be receptive to proactive interventions meant to counter the effects of heated or tense situations so that the discussion can be kept civil, productive, and generally on-track.

That said, proactive interventions are also much more informal than classical reactive moderation, and encompass a wide variety of "soft strategies" rather than any specific prescribed workflow. Therefore, an important starting point for any computational work on proactive interventions is to first characterize what these interventions look like in practice, in terms of what strategies are employed and when. Accordingly, this chapter explores in-depth what proactive interventions look like from the perspectives of both volunteer moderators and ordinary users. We achieve this by both synthesizing prior work on this topic and conducting our own qualitative interviews.

Note on source material. This chapter adapts and synthesizes material from Schluger et al. (2022) and Chang et al. (2022).

2.2 Background and Related Work

Two lines of prior work in computational social science and human-computer interaction are relevant to our goal of understanding the strategies underlying proactive interventions. One line of work focuses on the moderator perspective, examining the often-overlooked proactive steps that volunteer moderators take to maintain civility norms within their communities. Another line of work studies how technical interventions at the user interface level can be targeted

towards end-users to proactively discourage them from toxic behavior.

2.2.1 The Proactive Work of Volunteer Moderators

As noted in 1.1.2, volunteer moderators' dual identities, as both authority figures and regular community members, change the dynamics of their moderation work when compared to their professional, platform-employed counterparts. In interviews conducted by Seering et al. (2020), volunteer moderators described themselves as playing a diverse set of roles within their communities: on the one hand, they describe themselves as "mediators" or "police" in a nod to the fact that they are given substantive authority to enforce community rules much like professional moderators do, but on the other hand they also view themselves as "team members", "facilitators", "representatives", and "protectors"—metaphors that point to a sense of camaraderie and connection with their communities. Other work has used a parent-child metaphor as framing for this rather unique relationship (Shahid et al., 2024).

In addition to balancing multiple identities, volunteer moderators must simultaneously balance multiple possibly conflicting goals. As Grimmelmann (2015) observes, moderators generally hold openness as a core value for their communities—such that anyone can contribute freely—but they also tend to value productivity—which may require cracking down on certain kinds of contributions that impede collaboration and teamwork, such as toxicity. Likewise, Lo (2018)'s interviews with moderators reveal how they must regularly navigate keeping their communities regulated enough to remain safe and welcoming, but unregulated enough to feel spontaneous and fun.

This unique positioning of volunteer moderators within their communities simultaneously motivates and enables their efforts to proactively maintain social norms and prevent toxicity. Such proactive steps range in complexity and scale. At a broad, community-wide level, volunteer moderators may publicly model good behavior and conflict-resolution strategies in hopes that their fellow community members will follow suit (Jagannath et al., 2020; Seering et al., 2017), and broadcast general reminders about community rules (Cai and Wohn, 2019; Cai et al., 2021). At a more targeted level, volunteer moderators may choose to intervene in particularly sensitive or high-impact situations: for instance, personally going over the community's rules and norms with newcomers (Morgan and Halfaker, 2018; Seering et al., 2019b), fact-checking posts containing misinformation before they have a chance to spread and possibly cause conflict (Shahid et al., 2024), and mediating disputes that arise between different parties within the community (Billings and Watts, 2010).

Among the more targeted proactive intervention strategies, one in particular has recently picked up increasing interest: volunteer moderators actively monitor ongoing conversations in order to proactively prevent them from turning toxic or, at least, to be in a position that allows them to mitigate the effects of toxicity in a timely manner (Lo, 2018; Seering and Kairam, 2022). Unlike the more generic strategies discussed above, this requires substantial time and effort on behalf of the moderators due to the high volume of conversations they may need to actively monitor, and thus scales poorly. As such, recent work has advocated for offering algorithmic support for this strategy, proposing that predictive algorithms could be used to identify "at-risk" discussions that may be in need of monitoring, thereby helping moderators focus their finite attention where it is mostly likely to have an impact (Seering et al., 2019b; Jurgens et al.,

2019). Exploring this possibility comprises one core focus of our present work.

2.2.2 Well-intentioned Users and Conversational Derailment

As we have noted above, a key challenge for volunteer moderators seeking to proactively monitor at-risk discussions is knowing which discussions are at-risk. Moderators have reported that they tend to, over time, build up an intuition for when a currently civil discussion is in danger of turning toxic (Tiffany, 2019)—a key phenomenon we refer to as *conversational derailment*. Yet such intuition remains somewhat vague, leaving unclear the precise mechanisms by which conversations derail. To gain a deeper understanding, we must examine this phenomenon from the perspective of the users themselves: why is it that well-intentioned users, who are acting in good faith and do not explicitly seek to be toxic, can nevertheless end up acting toxic in initially-civil discussions?

An answer to this question begins to emerge from Cheng et al. (2017)'s pioneering study on antisocial behavior in online communities. Cheng et al. found that under the right circumstances, in their words, "anyone can become a troll". They specifically identified three types of factors that can lead even well-intentioned users to behave like trolls (i.e., in toxic ways). First, they point to the influence of offline factors: a user may simply be in a bad mood when posting a reply. Second, they find evidence that conversational context can also drive future toxicity, with properties like existing negative sentiment or high number of downvotes establishing an overall tone that is conducive to toxicity. Finally, they formalize a model for tracking the contagion of toxicity, showing that toxicity, left unchecked, can spread from user to user. Subsequent work

has provided further evidence for these findings; for instance, studies of online gaming communities, infamous for their toxic culture, have shown similar findings regarding the spread of toxicity through context and contagion (Kou, 2020; Shen et al., 2020).

Other work has shed light on additional factors that may contribute to well-intentioned users crossing the line into outright toxic behavior. Kumar et al. (2018) points to the negative influence of echo chambers, which can over time radicalize users to become more prone to conflict and confrontation, especially when interacting with other users from outside the echo chamber. McKee (2002) qualitatively argued that miscommunication between users can drive conflict that turns toxic, showing examples of forum posts that were intended to be innocuous but, for reasons ranging from ignorance to cultural gaps to simply bad wording choice, were (reasonably) perceived as inflammatory by other users. Our own work has subsequently found quantitative evidence for this miscommunication effect, finding that social media comments which were intended to share a fact but (mis)perceived as sharing an opinion are more likely to be followed by toxic replies (Chang et al., 2020a).

Promisingly, however, the fact that a well-intentioned user is led to toxicity once due to contextual or situational factors does not imply that they have been permanently transformed into a troll. Indeed, interviews with users who got penalized by moderators for toxic behavior reveal that many may express regret after the fact (Jhaver et al., 2019a). While this could cynically be interpreted as merely regret over being caught, our work has found evidence that this is not the case: in a study of the long-term outcomes of Wikipedia editors who were temporarily blocked for engaging in personal attacks against fellow editors, we

found that those who expressed remorse for their actions (e.g., in communications with the administrator who blocked them) were less likely to later have a repeated offense when compared to editors who instead expressed anger or frustration over "unfair" moderation, suggesting that such remorse is sincere (Chang and Danescu-Niculescu-Mizil, 2019a).

The latter finding also points to the possibility that carefully designed moderation policies can help discourage well-intentioned users from turning toxic, by fostering a sense that moderation is fair and on the side of the users. Subsequent work has found both qualitative and quantitative evidence to support this hypothesis: Shahid et al. (2024) finds that volunteer moderators in WhatsApp groups have found success by being more transparent about their moderation, while Weld et al. (2024) finds that when moderators are more engaged with their communities, users tend to have more positive sentiment towards moderation.

2.3 Methods

This chapter aims to build upon prior work on the dynamics of toxic behavior among well-intentioned users and effective strategies for managing it, by gleaning insights from interviews with moderators and end users alike. We conduct interviews in two settings whose userbases are reflective of the ideal of well-intentioned users: Wikipedia Talk Pages and the Reddit debate forum ChangeMyView.

2.3.1 Experimental Settings

Wikipedia Talk Pages

While Wikipedia is primarily known as an online encyclopedia, it also plays host to a vibrant community of editors who continually write new articles and improve existing ones. To support this community, Wikipedia has a feature known as *talk pages*: special pages on which editors can discuss a particular article or Wikipedia policy, or simply unwind with casual conversation. Every Wikipedia article has an associated talk page, on which editors can discuss proposed edits to the article (Kittur and Kraut, 2008).[1] In this collaborative, goal-driven discussion environment, toxicity is particularly impactful, threatening the health of the editor community and disrupting productivity (Henner and Sefidari, 2016; Kittur et al., 2007).

Moderation of Talk Page discussions is community driven (Seering, 2020): the Wikipedia community elects administrators with broad technical powers on the platform such as deleting articles or blocking other users.[2] A subset of these administrators choose to engage in discussion moderation. We note that there is no formal designation distinguishing discussion moderators from the rest of the administrators, and that discussion moderation practices (e.g., when a personal attack is subject for removal) are left largely at the discretion of these administrators.[3]

[1] See the Wikipedia talk page guidelines: `https://en.wikipedia.org/wiki/Wikipedia:Talk_page`

[2] `https://en.wikipedia.org/wiki/Wikipedia:Administrators`

[3] In addition, the community grants an elected committee of arbitrators even broader powers to impose binding resolutions in order to resolve particularly severe disputes on Wikipedia, including but not limited to disputes in discussions (`https://en.wikipedia.org/wiki/Wikipedia:Arbitration`).

Taken together, the goal-driven nature of Wikipedia Talk Page discussions and the large degree of discretion given to moderators make this a convenient setting for an initial case study of proactive moderation practices. Moreover, we believe the goal-driven nature of the discussions provides moderators with a strong motivation to improve their moderation practices, while the large degree of discretion granted to Wikipedia moderators gives them the freedom to consider and attempt alternative strategies. That said, it is important to note upfront that this setting also imposes some limitations on our work. Given the unique structure and culture of Wikipedia, our goal is not to report findings that generalize to any type of platform, but rather to begin understanding proactive moderation practices in the specific setting of goal-driven online discussions. In the process, we provide a blueprint that other researchers can follow to begin understanding proactive moderation in other types of online communities, both for its own sake and for comparison with this setting.

ChangeMyView

ChangeMyView[4] is a subreddit centered around good-faith debates, where the premise is that users come in with an opinion that they want to be challenged on, and invite other users to chime in with arguments that might convince them to alter or drop that opinion—that is, to "change their view". Given that the opinions users may bring to the table can sometimes be controversial, maintaining a culture of good-faith debates requires vigilant moderation, conducted (as is generally the case in subreddits) by volunteer moderators. To this end, the ChangeMyView moderators have over time developed a strict set of rules

[4] `https://www.reddit.com/r/changemyview`

governing what is and is not acceptable behavior[5]—including, most relevantly for our work, Rule 2 which forbids being "rude or hostile to other users". Similar to our reasoning about Wikipedia Talk Pages, we chose ChangeMyView as an ideal setting for our work because of this focus on good-faith interaction and strong moderation, as well as the fact that it has an established history of research collaborations (Jhaver et al., 2017; Hidey et al., 2017; Wei et al., 2016; Tan et al., 2016).

2.3.2 Interviews

Moderator Interviews

Following a rich line of prior literature that uses interviews to pull back the curtain on moderation practices (Gurzick et al., 2009; Dosono and Semaan, 2019; Wohn, 2019; Chandrasekharan et al., 2019; Seering et al., 2019b), we conducted semi-structured interviews with nine administrators on Wikipedia who engage in Talk Page moderation. Each interview was conducted over Zoom and lasted approximately one hour; we subsequently produced full de-identified transcripts and coded the data using thematic analysis. The interview questions asked participants for their thoughts on the role of administrators in moderation on Wikipedia, the goals of moderation, the ways they moderate proactively, and how they reason about the future of conversations to inform their proactive interventions. The generic script of the interviews is included in Appendix A.

We conducted these interviews with Institutional Review Board (IRB) approval and recruited participants through snowball sampling: by asking each

[5] https://www.reddit.com/r/changemyview/wiki/rules/

participant to recommend any other individuals they know who do discussion moderation on Wikipedia. While our interviews provided invaluable direct access to moderators and their domain specific knowledge, this recruitment procedure does impose a potential limitation on our work by potentially biasing our findings to the one branch of the Wikipedia moderator social graph that our sampling procedure reached.

User Surveys

Like the moderator interviews, our user interviews follow a rich line of work that uses surveys to examine ordinary social media users' experiences with toxicity and moderation (Jhaver et al., 2019a; Weld et al., 2024). Using an online survey, we asked 47 ChangeMyView users about their prior experiences with toxicity (defined as violations of Rule 2) and moderation, including what effects they think toxicity has on discussions, how they personally react to toxicity, and how effective moderation has been in their experience. Our use of an online survey, as opposed to Zoom interviews as in the case of moderators, was both for the sake of scalability and in recognition of Reddit's much stronger culture of anonymity. The survey was conducted with IRB approval, and the full text of the questions can be found, alongside further details about the execution of the study, in Appendix B.

2.4 Findings

2.4.1 Moderator Goals: Content and Environment

To contextualize our discussion, we start with the broad goals moderators have in our particular domain of Wikipedia Talk Page discussions. Following from the goal oriented nature of these discussions, as discussed in Section 2.3.1, participants highlighted how maintaining civil and productive discussions is not the end goal of their moderation. Rather, keeping discussions civil and functional is a crucial intermediary goal towards their primary goals: maintaining high quality content on the platform—in this case, encyclopedia articles—and maintaining a good environment for editors. As **PW6** puts it:

> **PW6**: [When I find a conversation headed downhill] I would not really care about the threads as having the thing go on, I'd care about the article and the environment of Wikipedia. I think those are the two things that I care about.

Discussion moderation is crucial to maintaining these goals: toxicity in discussions contributes directly to a hostile platform environment. Moreover, it can threaten the platform's content when it pushes editors to give up on editing an article or leave Wikipedia altogether (Wikimedia Support and Safety Team, 2015), or when it prevents or distracts from the conversations necessary for content creation and refinement. This finding corroborates prior work showing how volunteer moderators are motivated by, and must struggle to balance, a variety of goals, as discussed in Section 2.2.1.

A further consideration for moderators is that Wikipedia relies heavily on experienced users to contribute to the articles (Halfaker et al., 2013); when these important content creators act toxic in a discussion, moderators are hesitant to sanction them because of their perceived value in writing articles *even though* this incivility threatens the Wikipedia environment and alienates other users (Halfaker et al., 2011a; Collier and Bear, 2012). This exposes one way that the dual goals of moderation are in tension on Wikipedia. As **PW3** explains:

> **PW3**: I do believe that the English Wikipedia as a whole has a civility problem. [...] The community as a whole is far too willing to forgive incivility in the name of well—they're an experienced administrator or they're a really good content creator, so we'll just let them get by or say it wasn't that bad. And I think that that is not the path to a healthy community in the long term. I mean we have an editor retention problem, we know that. Everybody knows that. And I do think that the civility of the community is a significant part of that.

In their view, moderators' imbalanced approach to the dual goals of moderation threatens the platform overall, and contributes to the difficulty retaining users—illustrating the hard tradeoffs involved in balancing moderation goals.

2.4.2 Proactive Moderation Practices

Acting Proactively

Considering the broad goals of moderation on Wikipedia, we move to address one of our main research questions: Do moderators act proactively to prevent

discussions from derailing into toxicity, and if so, what is their workflow?

First, we confirm that moderators on Wikipedia do in fact engage in a variety of proactive moderation strategies. The starting point in their workflow is their ability to foresee whether a conversation is at risk of derailing. If they consider that this risk is elevated, they can further start monitoring it, or even decide to intervene in the discussion to avoid future derailment. For example:

> **PW6**: Sometimes I can sit by and see things developing and I might drop by with a comment. I don't tend to get involved in very big issues and charge in but I will go in and say, 'This is becoming an inappropriate way of speaking. Let's talk collaboratively. Let's talk constructively.' But do I monitor ongoing discussions for it? I suppose I look at some of the administrator notice boards, but I suppose I actually tend to sit more on the sidelines and watch other people engage in things, and only come in if I felt I had something to contribute or something to say like, 'Tone this down.' And there is a good chance somebody else might too.

While moderators have access to formal administrative tools, called sanctions on Wikipedia[6]—such as blocking and interaction bans—proactively imposing any formal sanction is not permitted by Wikipedia's moderation guidelines and would raise ethical concerns; sanctions can only be used in response to a tangible offense. Therefore, the proactive interventions that moderators can employ are limited to informal moderation techniques.

[6]From the Wikipedia:Sanctions page: "Sanctions are restrictions on editing Wikipedia that are applied to users or topic areas by the Wikipedia community and the Arbitration Committee in order to resolve disputes and curtail disruptive behaviour." (`https://en.wikipedia.org/wiki/Wikipedia:Sanctions`)

Participants identify a variety of informal strategies they use to guide conversations which they assess to be at risk of derailment. For example, moderators will join a discussion as a level-headed participant in order to refocus the discussion on its original topic. **PW5** explains their strategy:

> **PW5**: In some of those cases I just engage as an additional participant rather than in discussion moderation just in order to just try and aid in those methods by bringing the discussion back on to context.

A similar strategy is to leave just one comment in a discussion to acknowledge a growing dispute and try to neutralize it before it gets out of hand and irreparably damages the conversation. Prior work has described this as a moderator acting as a "mediator", stepping into a conversation facing rising tensions in order to resolve conflicts between clashing discussion participants (Seering et al., 2020). **PW8** explains their strategy:

> **PW8**: I'll just leave a comment being like, 'Hey guys, I think this might be going off topic,' and then I'll give my version of events. So it will be my opinion on it, in a very neutral way where I address each of their concerns. If I do it in a very polite way I think that typically a third party—especially an admin—does put the conversation back on topic.

A different version of this strategy is to remind users of platform rules when moderators anticipate they will be violated. Prior work has described this mode as a moderator acting as a "referee", working to "resolve disputes by referencing rules or a body of accepted knowledge" (Seering et al., 2020). This can be seen as

a more targeted version of automatic reminders, such as those triggered when interacting with newcomers (Halfaker et al., 2011b). **PW4** explains this strategy as they apply it:

> **PW4**: When we have [discussions in which there seems to be a significant chance of undesirable behavior arising], periodically we'll put up notices like, 'Hey, remember to keep civil, keep your comments about the content of the discussion, not the other editors directly.'

These three interventions show the wide range in the depth of moderator involvement required for different proactive interventions. Joining a discussion as a participant to try to bring it back on track requires contextual knowledge of the conversation topic at hand and continued involvement in a discussion. Similarly, leaving one comment to address the concerns of discussion participants requires contextual knowledge of the conversation and topic at hand, but does not require ongoing engagement. Finally, reminding users of the platform's policies only requires a prediction of which policies may be violated, while the reminder itself can take the same form across discussions on different topics and does not require continued engagement.

While some participants discuss how their proactive interventions can often bring discussions back on topic and avoid severe derailment beyond hope of repair, other participants describe how proactive interventions can backfire. Even when moderators forgo their formal sanctioning powers in favor of a softer approach, some users may react negatively to what they perceive as a threat of future sanctions. This implication may alienate users and limit the effectiveness of any proactive intervention. **PW1** explains:

> **PW1**: I did [proactive interventions in discussions] much more when I was younger. It doesn't work very well, I think because the idea is if you're coming in as sort of like an uninvolved administrator, [...] the assumed context is that you're getting ready to sanction them, which is never as useful as a friendly reminder. If I personally know one of the parties to the dispute, which happens on occasion, I might send them a direct email or a direct message, [...] just to try to hear what's going on. I found it particularly ineffective to post on Wiki to cool down, or something.

This highlights one specific challenge moderators face when acting proactively: demonstrating to users that they genuinely want to help the conversation remain on track and free of toxicity, rather than arriving early in preparation for future sanctions. This corroborates prior work showing how discussion moderators may shy away from joining discussions despite a desire to do so, because of their role as a moderator (Gurzick et al., 2009). Thus, executing a successful proactive intervention requires a nuanced approach that considers the ways a moderator's actions will be perceived by users.

Benefits of Acting Proactively

In addition to the established drawbacks of reactive moderation—and the respective benefits of the proactive paradigm—discussed in prior work and elaborated in Chapter 1, our interviews shed light on a further issue: echoing findings from prior work about conflicting goals in moderation (Section 2.2.1), reactive interventions struggle to balance the dual goals of ensuring high-quality content creation and maintaining a positive interactional environment (Section 2.4.1).

Since the reactive paradigm is only to act after a clear violation of community norms, in this case moderators can and do impose alienating formal sanctions. So, when experienced users who make otherwise valuable content engage in toxic behavior, actions to sanction them—intended to maintain a healthy environment on the platform—alienate them and hence threaten the further development of the platform. On the other hand, protecting these toxic users just because they create good content can cause disruption to the platform environment and alienate other users. **PW7** explains this conundrum:

> **PW7**: [When experienced editors clash,] that's where we, as administrators, sometimes have a very difficult task. We don't want to block experienced editors because they are very useful, very valuable. [...] By the same token, we don't want disruption. So, we've walked this very fine line where we try to hold experienced users who are sometimes misbehaving accountable without trying to block. It is a very difficult and fine line to walk and I think it would be nice if we had some way to better keep people civil, and better [...] get people to work together.

Thus, in the reactive paradigm, toxicity can threaten moderators' goals regardless of whether or not is addressed—disrupting the environment if it is not sanctioned, or alienating high value users if it is. Moreover, moderators face a significant challenge in realizing their dual moderation goals in the face of toxicity from established users through the reactive paradigm, threatening their emotional health and consuming a lot of their time. **PW2** explains:

> **PW2**: [When] someone has been incredibly uncivil to lots and lots of

> people, but he's also an incredibly influential editor, it is an excruciating process to kind of get through the kind of pieces that I need to to try and rein in his incivility. I just have to be patient, [because] it's ongoing and long.

Therefore, addressing toxicity from valuable content creators through the reactive paradigm threatens moderators themselves, in addition to their goals.

Where reactive moderation faces this dilemma, the proactive paradigm offers a solution. Because proactive interventions come before any tangible toxicity in a conversation, they are more well suited to take a softer and less alienating form. This allows moderators to support a healthy environment by preventing toxicity in discussions while avoiding the drawbacks of reactive strategies. **PW2** explains their preference for using the proactive paradigm to address rising tensions in a conversation:

> **PW2**: I did not become an administrator in order to block people. There are definitely people that became administrators because that's what they want to do, they wanted to police behavior. I actually spend a fair amount of time policing behavior in terms of my overall workload, but like I said, I try to operate in the social sphere and really kind of have conversations rather than using that.

While not all moderators share this preference, proactive moderation offers those who do use it a more nuanced approach to moderation, better suited to balance their multiple moderation goals, rather than appeal directly to one or the other.

Foreseeing Future Derailment

One crucial prerequisite of proactive moderation is identifying which conversations are at risk of turning toxic. We find that moderators on Wikipedia use their own intuition about the future trajectory of conversations towards this end, considering a variety of factors to internally reason about the future of the conversations they see. For example:

> Q: Given a civil conversation, do you think it is possible to predict if it will eventually derail into uncivil comments?
>
> **PW7**: Yes. Not always but yes. I would say, certainly with experience, you get a feel for it where if a discussion has started off on the wrong foot, maybe someone got [their edits] reverted and then they opened, you know, maybe not an uncivil but kind of a terse message like, "Hey, why did you undo my edit?," that's not uncivil but...It started things off on a bit of a wrong foot. I could guess that some of those things might get uncivil.

Moderators use a variety of factors to make predictions about the future of conversations. Five participants report using direct observations from the conversation, like the conversation content or tone, to do forecasting. Using these direct features allows moderators to update their predictions over time as the conversation develops, whenever they check in on the conversation. On the other hand, the other four participants report forecasting solely based on metadata, including features of the conversation and of the interlocutors. Salient conversation properties identified by participants include the ostensible conversation topic (as indicated by the conversation header) and the number of

participants in the conversation. Salient interlocutor metadata include level of experience on the platform, identity, and usernames. Drawing on their past experiences, participants consider such features to estimate the risk that a conversation is likely to derail in the future.

2.4.3 Evaluating the Feasibility of Algorithmic Assistance for Proactive Moderation

Equipped with an understanding of moderators' goals and practices, we now proceed to explore concrete ways in which an algorithmic tool can assist with their proactive moderation workflow. We consider components of the workflow where moderators suggest that technical support is needed, and assess the feasibility of offering this support algorithmically with existing technology in an ethical and efficient manner. From this discussion, we identify a potential role for algorithmic assistance in one crucial aspect of the workflow: discovering and monitoring at-risk conversations.

Discovery of At-risk Conversations: Need and Support

We previously uncovered how moderators use their own intuition to decide which conversations to proactively moderate; now, we turn to the challenges moderators face in this crucial process and the resulting need for additional support.

One idealized form of proactive moderation that all participants found appealing is to identify conversations that they suspect are highly likely to derail

and monitor them so that they can intervene proactively at an opportune moment or to react immediately to any uncivil behavior that does arise. However, moderators' ability to identify at-risk conversations to monitor is limited by the scale of the platform. **PW9** explains how even within the subset of topics they are interested in and engage in, their ability to effectively proactively monitor conversations is limited by their sheer number, which forces them to use only simplistic strategies, such as random discovery, to identify at-risk conversations to monitor:

> **PW9**: There are too many [potentially at-risk conversations] to proactively monitor. I know there's about 65 or 60 ongoing ones which are certainly always going to be at risk. [...] So I usually either wait until I'm asked, or I happen to see something, or I skip around and happen to notice something.

The problem of scale is exacerbated by the inherent difficulty of determining when a conversation is in need of a proactive intervention. While *every* participant we interviewed believes there are some contexts in which they can foresee derailment, as described in Section 2.4.2, there is a wide range in how broad this context is and how confident participants are in their forecasts. Four participants believe that they can confidently forecast antisocial behavior in any Wikipedia context, but four others believe that they can only do so in very specific contexts with low confidence, and the last participant believes they can only make such forecasts in conversations on a handful of specific topics among discussion participants they know personally.

Given that moderators are often uncertain about their forecasts of a conversation's future trajectory, they may hesitate to intervene immediately, and

instead desire to keep an eye on the situation to see how it develops. **PW3** explains:

> **PW3**: From time to time I do see a discussion I think that I want to monitor, and I'm like 'Yeah, I probably should be keeping an eye on this.' [...] I might leave a tab open on it and come back to it just in case.

As **PW3** goes on to elaborate, however, this idealized notion of monitoring a conversation as it develops in real time is impractical in reality:

> **PW3**: There are some technical challenges to [monitoring a discussion] just because of the way the Wikipedia software works. There isn't an easy way to say, 'Give me updates for any changes in this discussion.' And, in fact, you can't even say, 'Give me an update every time this page is changed,' which is a perennial source of annoyance.

But on the other hand, the resulting gap in attention could cause the moderator to miss out on key developments in the conversation, and thereby lose an opportunity to intervene. **PW6** explains this dilemma:

> **PW6**: I think I am okay at gauging if things are going to go pear-shaped, but do I always stick around to even find out if I am not interested in the topic? I may just move on and it blows up behind me. The hand grenade has gone off and I didn't even hear it because I've gone down the street.

We therefore find that proactive moderation practices are difficult to scale

up manually, both because of the size of the platform itself and because monitoring conversations—a necessary step given the uncertainty of moderators' forecasts—is time-consuming and impractical. We argue that these findings motivate work on algorithmic tools for proactive moderation: if an algorithm could accurately and efficiently identify conversations that are at risk, it could be use to automate the (currently manual and repetitive) process of monitoring conversations for relevant changes in derailment risk. This can potentially help moderators engage in proactive monitoring at a larger scale and dedicate more time to addressing potential issues.

2.4.4 Users' Experiences With Moderation

Having heard volunteer moderators' perspectives on toxicity, we now turn to the other side of the equation: what does moderation look like from the perspective of ordinary users? While our focus remains on proactive interventions, our conversations with users on the topic of moderation largely focused on classical reactive moderation, since this is the form of moderation that is most common and most visible to everyday users of ChangeMyView.

Participants' overall impression of the moderation of Rule 2 is somewhat mixed, with less than half of participants (46.8%) reporting that they are satisfied with enforcement of the rule. However, this dissatisfaction does not come from a place of disagreement with moderator actions: only 4.3% of participants considered enforcement of Rule 2 to be too strict, and only 6.4% considered it to be not strict enough. Rather, a chief source of dissatisfaction appears related to the fundamental drawback of reactive moderation: that it requires a moderator

to manually remove a toxic comment after-the-fact, while the toxic comment can continue to produce negative repercussions in the meantime. As **PR28**, **PR42**, and **PR46** describe,

> **PR28**: I wish [moderation] was faster, often somebody who is soapboxing will leave a string of uncivil comments and it stays up a while before they are removed.

> **PR42**: CMV removes certain comments, but far after the conversation dissolves into insults and hostility.

> **PR46**: Mainly [I would like to see] just faster enforcement. Enforcement after an hour is effectively useless.

In total, only 8.5% of participants reported that most uncivil comments they have encountered are immediately removed; by contrast, 53.2% reported that removal usually takes at least a few hours and 31.9% reported that removal usually takes a day or more.

The consequences of allowing toxic comments to stay up, even if only for a few hours, can be devastating from the user perspective. In cases where toxic comments were not immediately removed, it is highly unlikely for the conversation to recover on its own back to civil discourse: only 10.6% of participants report seeing this happen. Instead, the far more likely outcomes are that the conversation escalates into further toxicity (reported by 46.8% of participants) or simply dies out (reported by 38.3% of participants).

2.4.5 Users' Intuitions About Risk of Derailment

Given that reactive moderation (understandably) cannot prevent all cases of toxic outcomes, many users have had to adopt their own strategies for handling and avoiding toxicity. Chief among these is that *every* participant in our study reported that—much like moderators—they have at least some level of intuition for when a discussion is at risk of turning toxic. Explanations of how this intuition works vary across participants. Some participants reason about risk in terms of specific word choices:

> **PR7**: Referring to someone as "you" tends to signal things may take a turn, as does using generalizing language and absolute terms like "always" and "all".

> **PR17**: The easiest way is to analyze the phrasing. Stern, short phrases, completely contradicting the other person's viewpoint might come off as hostile and aggressive, causing a defensive reaction that might turn into an uncivil discussion.

Meanwhile, other participants look at higher-level concepts such as tone, and especially the sense that an interlocutor is making things personal:

> **PR34**: There is a certain tone or rhetorical posture that people will take prior or during an uncivil reply that forecasts their position. Often times folks that are uncivil also project a greater deal of certainty about their conclusions and will be quicker to disagree or criticize than they are to interrogate the position they disagree with.

PR33: The arguments diverge from the topic to trying to guess what the other is supposedly thinking or making assumptions about the person and try to associate them with groups/beliefs etc.

PR18: [The conversation might be at risk] if the conversation starts getting personal, attacking personal credentials or identity instead of the problem.

Then, most participants went on to report that this intuition shapes their subsequent behavior: 61.7% report that they are less likely to join a discussion they suspect to be at risk of derailing, and 76.6% say that if they do join they will change how they phrase their reply.

2.4.6 Users' Proactive Strategies for Handling At-risk Situations

Thus far, we have seen that not only do users generally claim to have intuition for when a conversation is at risk of turning toxic, a vast majority (over three-quarters) report that this intuition affects how they phrase their reply. This finding is especially promising, in that it suggests many well-intentioned users are willing to spend effort on proactively avoiding escalating tense situations. But how exactly do these users change the phrasing of their replies? To explore this question in a more focused way, we specifically consider a set of linguistic phenomena that have been connected to (in)civility and healthy interactions in prior work:

- **Politeness**: Linguists have long theorized that politeness serves as a buffer to soften the perceived force of a message (Brown and Levinson, 1987; Lakoff, 1973), and recent work has empirically validated this (Zhang et al., 2018a).
- **Formality**: Formality has been theorized to play a role in preventing misunderstanding (Heylighen and Dewaele, 1999), and in turn misunderstanding has been identified as a potential driver of incivility (Chang et al., 2020a).
- **Objectivity**: In the survey and in this work, we specifically define "objective" language as the use of facts and data in constructing a comment, in contrast with the use of personal experiences and emotions. This feature is more specific to our domain of ChangeMyView: we speculate that in the specific context of debates, reliance on fact-driven argumentation may help keep debates on topic and prevent descent into *ad hominems*, which may be connected to incivility (Habernal et al., 2018).
- **Question-asking**: Asking more questions might show an interest in engaging with the point of the interlocutor, and has previously been shown to prompt more positive feedback, such as liking and agreement, from interlocutors (Huang et al., 2017).
- **Swearing**: Swearing can be used to express aggression, but also to signal group identity or informality (Holgate et al., 2018).
- **Comment length**: In the context of debates, higher word count can indicate that the interlocutor is trying to be more explicit in their argument (O'Keefe, 1998, 1997), which may, like formality, reflect an attempt to avoid misunderstanding.

Out survey asks participants about their use of these strategies in conversations that they intuitively deem to be at risk. Among them, participants most commonly reported changes in four of them: increased politeness (52.7% of participants), use of more objective language (66.7%), asking more questions (50.0%), and use of more formal language (47.2%).

Additionally, we offered a free response "Other" option so participants could describe strategies that don't fit under any of the listed options; a quarter of participants took this option. Among these participants, many of the miscellaneous strategies mentioned reflect a theme of trying to avoid misunderstandings, which is consistent with the idea of miscommunication as a driver of derailment as identified by prior work (Section 2.2.2). For instance, **PR2** attempts to avoid misunderstandings by clarifying what the people in the conversation may have meant:

> **PR2**: I try to be extremely clear about what is, and what is not, being said or claimed (by either position). Often if the risk is due to a lack of clarity or miscommunication, such explicit clarification (of the question) can be helpful.

While similarly, **PR33** points to the idea of trying to establish common ground:

> **PR33**: The uncivility [sic] can come from a misunderstanding on the stances of the interlocutors. As such, clearly stating common ground and making calls to rationality can be a good tool to defuse a situation.

Overall, these findings show that well-intentioned users desire to avoid es-

calating at-risk conversations, and are willing to alter their behavior in order to achieve this goal. However, this does not imply that they are immune to engaging in toxic behavior themselves: 68.1% of participants actually report that they have at some point made a comment that they later regretted because in hindsight it was toxic. These regrettable actions may be driven by a number of factors. For one, sometimes users may be making an inaccurate judgment of risk; as **PR39** puts it:

> **PR39**: It's hard in the moment when reading a divisive comment to objectively recognize where the conversation is going.

There can also be uncertainty in judging how one's *own* contribution contributes to the risk, as **PR44** explains:

> **PR44**: I'm not always sure when what I'm going to say will make things better or worse.

Overall, 78.7% of participants expressed some degree of uncertainty about their risk intuitions, echoing **PR39** and **PR44**'s sentiments. These findings mirror our previous findings that proactively identifying at-risk situations is challenging for moderators as well (Section 2.4.3). As was the case in that setting, we speculate that these challenges represent a potential opening for algorithmic assistance: an additional nudge that enhances a user's awareness of existing tension in the conversation might support their existing efforts in preventing tense situations from escalating into outright toxic behavior.

2.5 Discussion

Motivated by the gap between the idealistic goal of preventing toxic behavior and the reality of existing reactive moderation strategies, this chapter has sought to deepen our understanding of the alternative proactive paradigm. Through a case study of moderation on Wikipedia Talk Pages, we uncover the workflow through which volunteer moderators proactively monitor and intervene in discussions they deem to be at risk of derailing into uncivil behavior. Similarly, through a survey of ChangeMyView users, we uncover how ordinary users fill in the gaps left by reactive moderation by intuitively predicting when a discussing might be at risk of turning toxic and carefully phrasing their replies to avoid such an outcome. In both cases, we identify challenges faced by moderators and users alike and argue for the potential of algorithmic assistance to meet these challenges.

From the moderator perspective, we reveal a delicate balance between two moderation goals in this collaborative setting: maintaining a civil and welcoming environment, while trying not to alienate otherwise valuable content creators. Reactive moderation tends to put these goals at odds: imposing harsh sanctions against toxic behavior from otherwise valuable contributors can alienate them, but leaving such behavior alone creates a less civil and less welcoming environment. Proactive moderation offers an alternative, by preventing sanctionable actions from occurring in the first place. Moreover, whereas reactive interventions tend to be strict formal sanctions such as a block, proactive interventions better lend themselves to more nuanced, informal actions. In interviews, moderators discuss how they employ proactive moderation strategies to prevent toxicity without needing to remove any content or alienate users.

From the user perspective, we reveal how the limitations of reactive moderation create a gap in which toxic comments may (temporarily) get a chance to negatively affect the outcomes of conversations, either by breeding further toxicity or by killing off potentially valuable discussions. In response to this gap, we have seen some ways that ordinary users take action to proactively prevent such outcomes: after identifying a conversation as being at risk of future toxicity, users may adjust their language to include more polite, objective, and formal language, and to generally reduce the likelihood of misunderstanding.

In both settings, our findings also reveal challenges faced by moderators and users alike in pursuing proactive strategies. Moderators reported that there are too many ongoing conversations for them to reasonably inspect, and that even when they manage to discover at-risk conversations, monitoring further developments in those conversations is logistically challenging. Similarly, users reported uncertainty in both their ability to identify at-risk conversations and their knowledge of what to do about it, which can lead them to inadvertently escalate the tension or even reply with a toxic comment they later regret. Inspired by these findings and by suggestions from prior work (Seering et al., 2019a; Jurgens et al., 2019), we conclude that these challenges provide initial motivation for developing algorithmic tools to assist in proactively identifying at-risk conversations. The fact that humans appear to have an intuition for when conversations are at-risk suggests that building such an algorithmic tool is at least a feasible goal. However, it is not immediately clear that existing technologies—which are generally optimized for after-the-fact, reactive detection of toxic content—are sufficient to power such a tool. Developing the algorithmic breakthroughs necessary for automatic detection of at-risk conversations, then, is the core technical challenge that sits at the center of the next few chapters.

CHAPTER 3

COMPUTATIONALLY EXPLORING CONVERSATIONAL DERAILMENT

3.1 Introduction

"Or vedi l'anime di color cui vinse l'ira."[1]

– Dante Alighieri, Divina Commedia, Inferno

Our findings in Chapter 2 have shed some light on the phenomenon of conversational derailment. Moderators and users alike report having some intuition for when a conversation is at risk of derailing—in other words, they are able to *forecast* derailment by paying attention to intuitive warning signs. This leads us to ask a natural question: is it possible to build algorithmic systems that can similarly forecast derailment?

We note that this goal is crucially different from that of prior computational work, which has focused on characterizing and detecting toxic behavior after the fact (Warner and Hirschberg, 2012; Davidson et al., 2017; Yin et al., 2009; Wulczyn et al., 2017; Chandrasekharan et al., 2017; Pavlopoulos et al., 2017b). Our goal, by contrast, is to detect *warning signs* indicating that a currently civil conversation is at risk of derailing into toxicity. Such warning signs could provide potentially actionable knowledge at a time when the conversation is still salvageable.

As a motivating example, consider the pair of conversations in Figure 3.1. Both exchanges took place in the context of the Wikipedia discussion page for

[1] "Now you see the souls of those whom anger overcame."

A1: Why there's no mention of it here? Namely, an altercation with a foreign intelligence group? True, by the standards of sources some require it wouln't even come close, not to mention having some really weak points, but it doesn't mean that it doesn't exist.	**B1:** Is the St. Petersberg Times considered a reliable source by wikipedia? It seems that the bulk of this article is coming from that one article, which speculates about missile launches and UFOs. I'm going to go through and try and find corroborating sources and maybe do a rewrite of the article. I don't think this article should rely on one so-so source.
A2: So what you're saying is we should put a bad source in the article because it exists?	**B2:** I would assume that it's as reliable as any other mainstream news source.

Figure 3.1: Two examples of initial exchanges from conversations concerning disagreements between editors working on the Wikipedia article about the Dyatlov Pass Incident. Only one of the conversations will eventually turn awry, with an interlocutor launching into a personal attack.

the article on the Dyatlov Pass Incident, and both show (ostensibly) civil disagreement between the participants. However, only one of these conversations will eventually derail into a personal attack ("Wow, you're coming off as a total d**k. [...] What the hell is wrong with you?"), while the other will remain civil.

As moderators and users noted in Chapter 2, humans have some intuition about which conversation is more likely to derail.[2] We may note the repeated, direct questioning with which **A1** opens the exchange, and that **A2** replies with yet another question. In contrast, **B1**'s softer, hedged approach ("it seems", "I don't think") appears to invite an exchange of ideas, and **B2** actually addresses the question instead of stonewalling. Could we endow artificial systems with such intuitions about the future trajectory of conversations?

In this chapter, we aim to computationally capture linguistic cues that pre-

[2]In fact, humans achieve an accuracy of 72% on this balanced task, showing that it is feasible, but far from trivial.

dict a conversation's future health. While we are not the first to computationally study conversational outcomes, most existing conversation modeling approaches aim to detect characteristics of an observed discussion or predict the outcome *after* the discussion concludes—e.g., whether it involves a present dispute (Allen et al., 2014; Wang and Cardie, 2014). Our goal, by contrast, is subtly different: we aim to discover interactional signals of the *future* trajectory of an *ongoing* conversation. Such a goal recognizes derailment as emerging from the development of the conversation, and belongs to the broader area of *conversational forecasting*, which includes future-prediction tasks such as predicting the eventual length of a conversation (Backstrom et al., 2013), whether a negotiation (Sokolova et al., 2008) or persuasion attempt (Tan et al., 2016; Wachsmuth et al., 2018; Yang et al., 2019) will eventually succeed, whether team discussions will eventually lead to an increase in performance (Niculae and Danescu-Niculescu-Mizil, 2016), or whether ongoing counseling conversations will eventually be perceived as helpful (Althoff et al., 2016). In pursuing this goal, we are the first to frame the problem of toxicity through the lens of conversational forecasting, thereby introducing a novel forecasting task.

We make a first approach to this problem by analyzing the role of politeness (or lack thereof) in keeping conversations on track. Prior work has shown that politeness can help shape the course of offline (Clark, 1979; Clark and Schunk, 1980), as well as online interactions (Burke and Kraut, 2008), through mechanisms such as softening the perceived force of a message (Fraser, 1980), acting as a buffer between conflicting interlocutor goals (Brown and Levinson, 1987), and enabling all parties to save face (Goffman, 1955). This suggests the potential of politeness to serve as an indicator of whether a conversation will sustain its initial civility or eventually derail.

Recent studies have computationally operationalized prior formulations of politeness by extracting linguistic cues that reflect politeness strategies (Danescu-Niculescu-Mizil et al., 2013a; Aubakirova and Bansal, 2016). Such research has additionally tied politeness to social factors such as individual status (Danescu-Niculescu-Mizil et al., 2012; Krishnan and Eisenstein, 2015), and the success of requests (Althoff et al., 2014) or of collaborative projects (Ortu et al., 2015). However, to the best of our knowledge, this is the first computational investigation of the relation between politeness strategies and the future trajectory of the conversations in which they are deployed. Furthermore, we generalize beyond predefined politeness strategies by using an unsupervised method to discover additional rhetorical prompts used to initiate different types of conversations that may be specific to online collaborative settings, such as coordinating work (Kittur and Kraut, 2008) or conducting factual checks.

We explore the role of such pragmatic and rhetorical devices in forecasting derailment of conversations between Wikipedia editors. For this purpose, we introduce a new dataset of Wikipedia Talk Page discussions, which we compile through a combination of machine learning and crowdsourced filtering. The dataset consists of conversations which begin with ostensibly civil comments, and either remain healthy or derail into personal attacks. Starting from this data, we construct a setting that mitigates effects which may trivialize the task. In particular, some topical contexts (such as politics and religion) are naturally more susceptible to antisocial behavior (Kittur et al., 2009; Cheng et al., 2015). We employ techniques from causal inference (Rosenbaum, 2010) to establish a controlled framework that focuses our study on topic-agnostic linguistic cues.

In this controlled setting, we find that pragmatic cues extracted from the very

first exchange in a conversation (i.e., the first comment-reply pair) can indeed provide some signal of whether the conversation will subsequently derail. For example, conversations prompted by hedged remarks sustain their initial civility more so than those prompted by forceful questions, or by direct language addressing the other interlocutor.

In summary, this chapter's main contributions are:

- We articulate the new task of forecasting whether an ongoing conversation will derail into personal attacks;
- We devise a controlled setting and build a labeled dataset to study this phenomenon;
- We investigate how politeness strategies and other rhetorical devices are tied to the future trajectory of a conversation.

More broadly, we show the feasibility of automatically detecting warning signs of future misbehavior in collaborative interactions. By providing a labeled dataset together with basic methodology and several baselines, we open the door to further work on understanding factors which may derail or sustain healthy online conversations. To facilitate such future explorations, we distribute the data and code as part of the open source ConvoKit Python package (Chang et al., 2020b).

Note on source material. This chapter was originally published as Zhang et al. (2018a). Since the original publication date, the `spaCy` Python package, which we use for dependency parsing in order to extract several key features, has updated its dependency parser algorithm several times leading to fluctuations in

the extracted feature values, with downstream effects on the final results. Consequently, the results reported here differ slightly from the ones in original publication. The current results are fully reproducible since they make use of the public ConvoKit code and data (which is distributed with a fixed set of dependency parses). In addition to the updated results, some minor changes have been made to the wording and naming of phenomena throughout, for the sake of consistency with other chapters in this thesis.

3.2 Further Related Work

Antisocial behavior. Prior work has studied a wide range of disruptive interactions in various online platforms like Reddit and Wikipedia, examining behaviors like aggression (Kayany, 1998), harassment (Chatzakou et al., 2017; Vitak et al., 2017), and bullying (Akbulut et al., 2010; Kwak et al., 2015; Singh et al., 2017), as well as their impact on aspects of engagement like user retention (Collier and Bear, 2012; Wikimedia Support and Safety Team, 2015) or discussion quality (Arazy et al., 2013). Several studies have sought to develop machine learning techniques to detect signatures of online toxicity, such as personal insults (Yin et al., 2009), harassment (Sood et al., 2012) and abusive language (Nobata et al., 2016; Gambäck and Sikdar, 2017; Pavlopoulos et al., 2017a; Wulczyn et al., 2017). These works focus on detecting toxic behavior after it has already occurred; a notable exception is (Cheng et al., 2017), which predicts future community enforcement against users in news-based discussions. Our work similarly aims to understand *future* toxicity; however, our focus is on studying the trajectory of a conversation rather than the behavior of individuals across disparate discussions.

Discourse analysis. Our present study builds on a large body of prior work in computationally modeling discourse. Both unsupervised (Ritter et al., 2010) and supervised (Zhang et al., 2017a) approaches have been used to categorize behavioral patterns on the basis of the language that ensues in a conversation, in the particular realm of online discussions. Models of conversational behavior have also been used to predict conversation outcomes, such as betrayal in games (Niculae et al., 2015), and success in team problem solving settings (Fu et al., 2017) or in persuading others (Tan et al., 2016; Zhang et al., 2016).

While we are inspired by the techniques employed in these approaches, our work is concerned with predicting the future trajectory of an ongoing conversation as opposed to a post-hoc outcome. In this sense, we build on prior work in modeling conversation trajectory, which has largely considered *structural* aspects of the conversation (Kumar et al., 2010; Backstrom et al., 2013). We complement these structural models by seeking to extract potential signals of future outcomes from the *linguistic discourse* within the conversation.

3.3 Finding Conversations That Derail

We develop our framework for understanding linguistic markers of conversational trajectories in the context of Wikipedia's *talk page* discussions—public forums in which contributors convene to deliberate on editing matters such as evaluating the quality of an article and reviewing the compliance of contributions with community guidelines. The dynamic of conversational derailment is particularly intriguing and consequential in this setting by virtue of its collaborative, goal-oriented nature. In contrast to unstructured commenting forums,

Job 1: Ends in personal attack. We show three annotators a conversation and ask them to determine if its last comment is a personal attack toward someone else in the conversation.		**Job 2: Civil start.** We split conversations into snippets of three consecutive comments. We ask three annotators to determine whether any of the comments in a snippet is toxic.	
Annotators	367	**Annotators**	247
Conversations	4,022	**Conversations**	1,252
Agreement	67.8%	**Snippets**	2,181
		Agreement	87.5%

Table 3.1: Descriptions of crowdsourcing jobs, with relevant statistics. More details in Appendix C.

cases where one *collaborator* turns on another over the course of an initially civil exchange constitute perplexing pathologies. In turn, these toxic attacks are especially disruptive in Wikipedia since they undermine the social fabric of the community as well as the ability of editors to contribute (Henner and Sefidari, 2016). To approach this domain we start from the WikiConv dataset, which contains roughly 50 million conversations across 16 million Talk Pages, constructed by translating sequences of revisions of each talk page into structured conversations (Hua et al., 2018).

Roughly one percent of Wikipedia comments are estimated to exhibit antisocial behavior (Wulczyn et al., 2017). This illustrates a challenge for studying conversational failure: one has to sift through many conversations in order to find even a small set of examples. To avoid such a prohibitively exhaustive analysis, we first use a machine learning classifier to identify candidate conversations that are likely to contain a toxic contribution, and then use crowdsourcing to vet the resulting labels and construct our controlled dataset.

Candidate selection. Our goal is to analyze how the start of a *civil* conversation

is tied to its potential future derailment into personal attacks. Thus, we only consider conversations that start out as ostensibly civil—i.e., where at least the first exchange does not exhibit any toxic behavior[3]—and that continue beyond this first exchange. To focus on the especially perplexing cases when the attacks come *from within*, we seek examples where the attack is initiated by one of the two participants in the initial exchange.

To select candidate conversations to include in our collection, we use the toxicity classifier provided by the Perspective API,[4] which is trained on Wikipedia talk page comments that have been annotated by crowdworkers (Thain et al., 2017). This provides a toxicity score t for all comments in our dataset, which we use to preselect two sets of conversations: (a) candidate conversations that are civil throughout, i.e., conversations in which all comments (including the initial exchange) are not labeled as toxic ($t < 0.4$); and (b) candidate conversations that turn toxic after the first (civil) exchange, i.e., conversations in which the N-th comment ($N > 2$) is labeled toxic ($t \geq 0.6$), but all the preceding comments are not ($t < 0.4$).

Crowdsourced filtering. Starting from these candidate sets, we use crowdsourcing to vet each conversation and select a subset that are perceived by humans to either stay civil throughout ("on-track" conversations), or start civil but end with a *personal attack* ("derailing" conversations). To inform the design of this human-filtering process and to check its effectiveness, we start from a seed set of 232 conversations manually verified by the authors to end in personal attacks (more details about the selection of the seed set and its role in the crowdsourcing process can be found in Appendix C). We take particular care to not

[3]For the sake of generality, in this work we focus on this most basic conversational unit: the first comment-reply pair starting a conversation.

[4]`https://www.perspectiveapi.com/`

over-constrain crowdworker interpretations of what personal attacks may be, and to separate toxicity from civil disagreement, which is recognized as a key aspect of effective collaborations (Coser, 1956; De Dreu and Weingart, 2003).

We design and deploy two filtering jobs using the CrowdFlower platform, summarized in Table 3.1 and detailed in Appendix C. **Job 1** is designed to select conversations that contain a "rude, insulting, or disrespectful" comment towards another user in the conversation—i.e., a personal attack. In contrast to prior work labeling antisocial comments in isolation (Sood et al., 2012; Wulczyn et al., 2017), annotators are asked to label personal attacks in the *context* of the conversations in which they occur, since antisocial behavior can often be context-dependent (Cheng et al., 2017). In fact, in order to ensure that the crowdworkers read the entire conversation, we also ask them to indicate who is the target of the attack. We apply this task to the set of candidate awry-turning conversations, selecting the 14% which all three annotators perceived as ending in a personal attack.[5]

Job 2 is designed to filter out conversations that do not actually start out as civil. We run this job to ensure that the *derailing* conversations are civil up to the point of the attack—i.e., actually *derail* from civil into toxic behavior—discarding 5% of the candidates that passed Job 1. We also use it to verify that the candidate *on-track* conversations are indeed civil throughout, discarding 1% of the respective candidates. In both cases we filter out conversations in which three annotators could identify at least one comment that is "rude, insulting, or disrespectful".

Controlled setting. Finally, we need to construct a setting that affords for mean-

[5]We opted to use unanimity in this task to account for the highly subjective nature of the phenomenon.

ingful comparison between conversations that derail and those that stay on track, and that accounts for trivial topical confounds (Kittur et al., 2009; Cheng et al., 2015). We mitigate topical confounds using matching, a technique developed for causal inference in observational studies (Rubin, 2007). Specifically, starting from our human-vetted collection of conversations, we pair each *derailing* conversation, with an *on-track* conversation, such that both took place on the same talk page. If we find multiple such pairs, we only keep the one in which the paired conversations take place closest in time, to tighten the control for topic. Conversations that cannot be paired are discarded.

This procedure yields a total of 1,168 paired derailing and on-track conversations (including our initial seed set), spanning 536 distinct talk pages (averaging 1.1 pairs per page, maximum 5). The average length of a conversation is 4.4 comments.

3.4 Capturing Pragmatic Devices

We now describe our framework for capturing linguistic cues that might inform a conversation's future trajectory. Crucially, given our focus on conversations that start seemingly civil, we do not expect overtly hostile language—such as insults (Yin et al., 2009)—to be informative. Instead, we seek to identify pragmatic markers within the initial exchange of a conversation that might serve to reveal or exacerbate underlying tensions that eventually come to the fore, or conversely suggest sustainable civility. In particular, in this work we explore how politeness strategies and rhetorical prompts reflect the future health of a conversation.

Prompt Type	Description	Example
`Factual check`	Statements about article content, pertaining to or contending issues like factual accuracy.	The census **is not talking about** families here.
`Moderation`	Rebukes or disputes concerning moderation decisions such as blocks and reversions.	**If** you continue, you may **be blocked** from editing.
`Coordination`	Requests, questions, and statements of intent.	It's a long list so I **could do with** your **help**.
`Casual remark`	Casual, highly conversational aside-remarks.	**What's with** this flag image?
`Action statement`	Requests, statements, and explanations about various editing actions.	**Please consider improving** the article to address the issues [...]
`Procedures`	Statements of Wikipedia editing policies which are not directly related to moderation.	Consider **verifying** that you have **specified sources** for those files [...]

Table 3.2: Prompt types automatically extracted from talk page conversations, with interpretations and examples from the data. Bolded text indicate common prompt phrasings extracted by the framework. Further examples are shown in Appendix D, Table D.1.

Politeness strategies. Politeness can reflect a-priori good will and help navigate potentially face-threatening acts (Goffman, 1955; Lakoff, 1973), and also offers hints to the underlying intentions of the interlocutors (Fraser, 1980). Hence, we may naturally expect certain politeness strategies to signal that a conversation is likely to stay on track, while others might signal derailment.

In particular, we consider a set of pragmatic devices signaling politeness drawn from (Brown and Levinson, 1987). These linguistic features reflect two overarching types of politeness. *Positive* politeness strategies encourage so-

cial connection and rapport, perhaps serving to maintain cohesion throughout a conversation; such strategies include gratitude ("*thanks* for your help"), greetings ("*hey*, how is your day so far") and use of "please", both at the start ("*Please* find sources for your edit...") and in the middle ("Could you *please* help with...?") of a sentence. *Negative* politeness strategies serve to dampen an interlocutor's imposition on an addressee, often through conveying indirectness or uncertainty on the part of the commenter. Both commenters in example **B** (Fig. 3.1) employ one such strategy, hedging, perhaps seeking to soften an impending disagreement about a source's reliability ("I *don't think*...", "I would *assume*..."). We also consider markers of *impolite* behavior, such as the use of direct questions ("*Why*'s there no mention of it?') and sentence-initial second person pronouns ("*Your* sources don't matter..."), which may serve as forceful-sounding contrasts to negative politeness markers. Following (Danescu-Niculescu-Mizil et al., 2013a), we extract such strategies by pattern matching on the dependency parses of comments.

Types of conversation prompts. To complement our pre-defined set of politeness strategies, we seek to capture domain-specific rhetorical patterns used to initiate conversations. For instance, in a collaborative setting, we may expect conversations that start with an invitation for working together to signal less tension between the participants than those that start with statements of dispute. We discover types of such *conversation prompts* in an unsupervised fashion by extending a framework used to infer the rhetorical role of questions in (offline) political debates (Zhang et al., 2017b) to more generally extract rhetorical roles of comments. The procedure follows the intuition that a comment's rhetorical role is reflected in the type of replies it is likely to elicit. As such, comments which tend to trigger similar replies constitute a particular type of prompt.

To implement this intuition, we derive two different low-rank representations of the common lexical phrasings contained in comments (agnostic to the particular topical content discussed), automatically extracted as recurring sets of arcs in the dependency parses of comments. First, we derive *reply-vectors* of phrasings, which reflect their propensities to *co-occur*. In particular, we perform singular value decomposition on a term-document matrix $\mathcal{R}$ of phrasings and replies as $\mathcal{R} \approx \hat{\mathcal{R}} = U_R S V_R^T$, where rows of U_R are low-rank reply-vectors for each phrasing.

Next, we derive *prompt-vectors* for the phrasings, which reflect similarities in the subsequent replies that a phrasing *prompts*. We construct a prompt-reply matrix $\mathcal{P} = (p_{ij})$ where $p_{ij} = 1$ if phrasing j occurred in a reply to a comment containing phrasing i. We project $\mathcal{P}$ into the same space as U_R by solving for $\hat{\mathcal{P}}$ in $\mathcal{P} = \hat{\mathcal{P}} S V_R^T$ as $\hat{\mathcal{P}} = \mathcal{P} V_R S^{-1}$. Each row of $\hat{\mathcal{P}}$ is then a prompt-vector of a phrasing, such that the prompt-vector for phrasing i is close to the reply-vector for phrasing j if comments with phrasing i tend to prompt replies with phrasing j. Clustering the rows of $\hat{\mathcal{P}}$ then yields k conversational *prompt types* that are unified by their similarity in the space of replies. To infer the prompt type of a new comment, we represent the comment as an average of the representations of its constituent phrasings (i.e., rows of $\hat{\mathcal{P}}$) and assign the resultant vector to a cluster.[6]

To determine the prompt types of comments in our dataset, we first apply the above procedure to derive a set of prompt types from a *disjoint* (unlabeled) corpus of Wikipedia talk page conversations (Danescu-Niculescu-Mizil et al., 2012). After initial examination of the framework's output on this external data,

[6]We scale rows of U_R and $\hat{\mathcal{P}}$ to unit norm. We assign comments whose vector representation has (ℓ_2) distance ≥ 1 to all cluster centroids to an extra, infrequently-occurring null type which we ignore in subsequent analyses.

we chose to extract $k = 6$ prompt types, shown in Table 3.2 along with our interpretations.[7] These prompts represent signatures of conversation-starters spanning a wide range of topics and contexts which reflect core elements of Wikipedia, such as moderation disputes and coordination (Kittur et al., 2007; Kittur and Kraut, 2008). We assign each comment in our present dataset to one of these types.[8]

3.5 Analysis

We are now equipped to computationally explore how the pragmatic devices used to start a conversation can signal its future health. Concretely, to quantify the relative propensity of a linguistic marker to occur at the start of derailing versus on-track conversations, we compute the log-odds ratio of the marker occurring in the initial exchange—i.e., in the first or second comments—of derailing conversations, compared to initial exchanges in the on-track setting. These quantities are depicted in Figure 3.2A.[9]

Focusing on the **first** comment (represented as ◊s), we find a rough correspondence between linguistic *directness* and the likelihood of future personal attacks. In particular, comments which contain *direct questions*, or exhibit *sentence-initial you* (i.e., "2nd person start"), tend to start derailing conversations signif-

[7] We experimented with more prompt types as well, finding that while the methodology recovered finer-grained types, and obtained qualitatively similar results and prediction accuracies as described in Sections 3.5 and 3.6, the assignment of comments to types was relatively sparse due to the small data size, resulting in a loss of statistical power.

[8] While the particular prompt types we discover are specific to Wikipedia, the methodology for inferring them is unsupervised and is applicable in other conversational settings.

[9] To reduce clutter we only depict features which occur a minimum of 50 times and have absolute log-odds ≥ 0.2 in at least one of the data subsets.

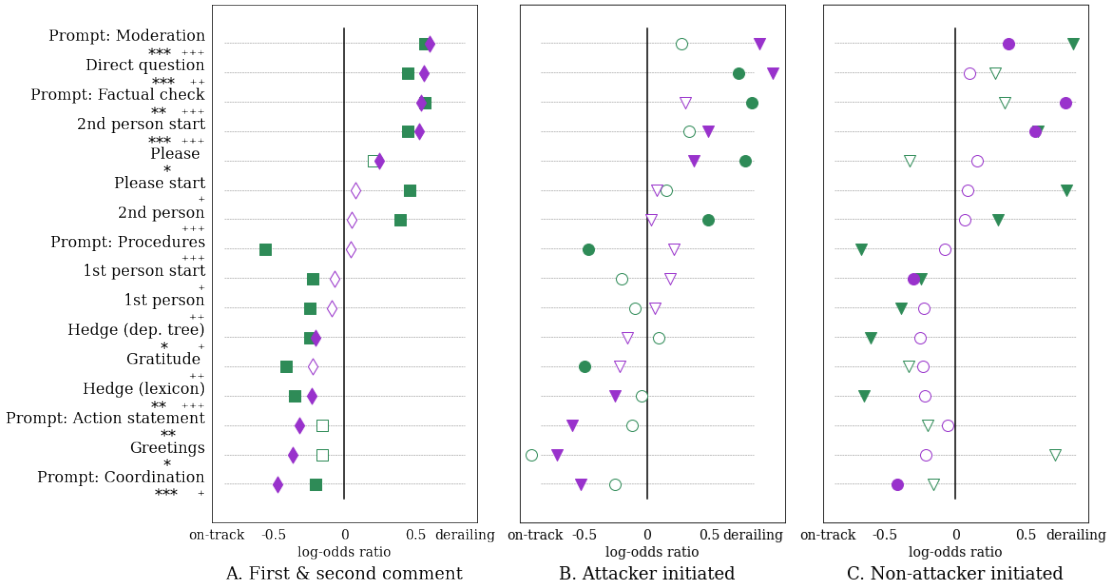


Figure 3.2: Log-odds ratios of politeness strategies and prompt types exhibited in the first and second comments of conversations that derail, versus those that stay on-track. **All**: Purple and green markers denote log-odds ratios in the first and second comments, respectively; points are solid if they reflect significant ($p < 0.05$) log-odds ratios with an effect size of at least 0.2. **A**: ♢s and □s denote **first** and **second** comment log-odds ratios, respectively; * denotes statistically significant differences at the $p < 0.05$ (*), $p < 0.01$ (**) and $p < 0.001$ (***) levels for the first comment (two-tailed binomial test); + denotes corresponding statistical significance for the second comment. **B** and **C**: ▽s and ◯s correspond to effect sizes in the comments authored by the **attacker** and **non-attacker**, respectively, in **attacker initiated** (**B**) and **non-attacker initiated** (**C**) conversations.

icantly more often than ones that stay on track (both $p < 0.001$).[10] This effect coheres with our intuition that directness signals some latent hostility from the conversation's initiator, and perhaps reinforces the forcefulness of contentious impositions (Brown and Levinson, 1987). This interpretation is also suggested by the relative propensity of the `factual check` prompt, which tends to cue disputes regarding an article's factual content ($p < 0.05$).

In contrast, comments which initiate on-track conversations tend to contain

[10] All p values in this section are computed as two-tailed binomial tests, comparing the proportion of derailing conversations exhibiting a particular device to the proportion of on-track conversations.

greetings ($p < 0.05$), an example of a positive politeness strategy. Such conversations are also more likely to begin with `coordination` ($p < 0.001$) and `action statement` ($p < 0.01$) prompts, signaling active efforts to foster constructive teamwork. Negative politeness strategies are salient in on-track conversations as well, reflected by the use of *hedges* ($p < 0.01$), which may serve to soften impositions or factual contentions (Hübler, 1983).

These effects are echoed in the **second** comment—i.e., the **first reply** (represented as □s). Interestingly, in this case we note that the difference in pronoun use is especially marked. First replies in conversations that eventually derail tend to contain more *second person pronouns* ($p < 0.001$), perhaps signifying a replier pushing back to contest the initiator; in contrast, on-track conversations have more *first person pronouns* ($p < 0.01$), potentially indicating the replier's willingness to step into the conversation and work with—rather than argue against—the initiator (Tausczik and Pennebaker, 2010).

Distinguishing interlocutor behaviors. Are the linguistic signals we observe solely driven by the eventual attacker, or do they reflect the behavior of both actors? To disentangle the attacker and non-attackers' roles in the initial exchange, we examine their language use in these two possible cases: when the *future* attacker initiates the conversation, or is the first to reply. In **attacker-initiated** conversations (Figure 3.2B, 558 conversations), we see that both actors exhibit a propensity for the linguistically direct markers (e.g., *direct questions*) that tend to signal future attacks. Some of these markers are used particularly often by the **non-attacking replier** in awry-turning conversations (e.g., *second person pronouns*, $p < 0.001$, ○s), further suggesting the dynamic of the replier pushing back at—and perhaps even escalating—the attacker's initial hint of ag-

gression. Among conversations initiated instead by the **non-attacker** (Figure 3.2C, 610 conversations), the non-attacker's linguistic behavior in the first comment (◯s) is less distinctive from that of initiators in the on-track setting (i.e., log-odds ratios closer to 0); markers of future derailment are (unsurprisingly) more pronounced once the eventual attacker (▽s) joins the conversation in the second comment.[11]

More broadly, these results reveal how different politeness strategies and rhetorical prompts deployed in the initial stages of a conversation are tied to its future trajectory.

3.6 Predicting Future Attacks

We now show that it is indeed feasible to predict whether a conversation will derail based on linguistic properties of its very first exchange, providing several baselines for this new task. In doing so, we demonstrate that the pragmatic devices examined above encode signals about the future trajectory of conversations, capturing some of the intuition humans are shown to have.

We consider the following balanced prediction task: given a pair of conversations, which one will eventually lead to a personal attack? We extract all features from the very first exchange in a conversation—i.e., a comment-reply pair, like those illustrated in our introductory example (Figure 3.1). We use logistic regression and report accuracies on a leave-one-page-out cross validation, such that in each fold, all conversation pairs from a given talk page are held out as

[11] As an interesting avenue for future work, we note that some markers used by non-attacking initiators potentially still anticipate later attacks, suggested by, e.g., the relative prevalence of `factual checks` ($p < 0.001$, ◯s).

test data and pairs from all other pages are used as training data (thus preventing the use of page-specific information). Prediction results are summarized in Table 3.3.

Language baselines. As baselines, we consider several straightforward features: word count (which performs at chance level), sentiment lexicon (Liu et al., 2005) and bag of words.

Pragmatic features. Next, we test the predictive power of the **prompt types** and **politeness strategies** features introduced in Section 3.4. The 12 prompt type features (6 features for each comment in the initial exchange) achieve 62.0% accuracy, and the 38 politeness strategies features (19 per comment) achieve 55.1% accuracy. These **pragmatic** features combine to reach 62.7% accuracy.

Reference points. To better contextualize the performance of our features, we compare their predictive accuracy to the following reference points:

Interlocutor features: Certain kinds of interlocutors are potentially more likely to be involved in derailing conversations. For example, perhaps newcomers or anonymous participants are more likely to derail interactions than more experienced editors. We consider a set of features representing participants' experience on Wikipedia (i.e., number of edits) and whether the comment authors are anonymous. In our task, these features perform at the level of random chance.

Trained toxicity: We also compare with the toxicity score of the exchange from the Perspective API classifier—a perhaps unfair reference point, since this supervised system was trained on additional human-labeled training examples from the same domain and since it was used to create the very data on which we evaluate. This results in an accuracy of 58.2%; combining trained toxicity

Feature set	# features	Accuracy
Bag of words	5,000	56.7%
Sentiment lexicon	4	55.4%
Politeness strategies	38	55.1%
Prompt types	12	62.0%
Pragmatic (all)	50	62.7%
Interlocutor features	5	51.2%
Trained toxicity	2	58.2%
*Toxicity + **Pragmatic***	52	66.0%
Humans		72.0%

Table 3.3: Accuracies for the balanced future-prediction task. Features based on pragmatic devices are **bolded**, reference points are *italicized*.

with our pragmatic features achieves 66.0%.

Humans: A sample of 100 pairs were labeled by (non-author) volunteer human annotators. They were asked to guess, from the initial exchange, which conversation in a pair will lead to a personal attack. Majority vote across three annotators was used to determine the human labels, resulting in an accuracy of 72%. This confirms that humans have some intuition about whether a conversation might be heading in a bad direction, which our features can partially capture. In fact, the classifier using pragmatic features is accurate on 80% of the examples that humans also got right.

Overall, these initial results show the feasibility of reconstructing some of the human intuition about the future trajectory of an ostensibly civil conversation in order to predict whether it will eventually turn awry.

3.7 Conclusions and Discussion

In this chapter, we started to examine the intriguing phenomenon of conversational derailment, studying how the use of pragmatic and rhetorical devices relates to future conversational derailment. Our investigation centers on the particularly perplexing scenario in which one participant of a civil discussion later attacks another, and explores the new task of predicting whether an initially healthy conversation will derail into such an attack. To this end, we develop a computational framework for analyzing how general politeness strategies and domain-specific rhetorical prompts deployed in the initial stages of a conversation are tied to its future trajectory.

Making use of machine learning and crowdsourcing tools, we formulate a tightly-controlled setting that enables us to meaningfully compare conversations that stay on track with those that derail. Human accuracy at forecasting derailment in this setting (72%) suggests it is feasible at least at the level of human intuition. We show that our computational framework can recover some of that intuition, hinting at the potential of automated methods to identify signals of the future trajectories of online conversations. We position this as a novel conversational forecasting task, for which these results offer a first baseline.

That said, our current approach is more a proof-of-concept than a practically useful solution for forecasting derailment, coming with several limitations that open avenues for future work. Our correlational analyses do not provide any insights into *causal* mechanisms of derailment, which randomized experiments could address. Additionally, since our procedure for collecting and vetting data focused on precision rather than recall, it might miss more subtle attacks that

are overlooked by the toxicity classifier. Supplementing our investigation with other indicators of antisocial behavior, such as editors blocking one another, could enrich the range of attacks we study. Noting that our framework is not specifically tied to Wikipedia, it would also be valuable to explore the varied ways in which this phenomenon arises in other (possibly non-collaborative) public discussion venues, such as Reddit and Facebook Pages.

Perhaps most significantly, while our analysis focused on the very first exchange in a conversation for the sake of generality, in reality signals of future derailment might be found throughout the conversation. Practically useful forecasting therefore requires more complex modeling that can account for such features that span an entire interaction. One promising path for achieving this is to go beyond the present binary classification task and explore a sequential formulation predicting whether the next turn is likely to be an attack as a discussion unfolds, capturing conversational dynamics such as sustained escalation. Implementing this idea, and addressing its associated technical challenges, is a key step towards practical algorithms for proactive interventions, which we explore in the next chapter.

CHAPTER 4

PRACTICAL FORECASTING OF CONVERSATIONAL DERAILMENT

4.1 Introduction

"Ché saetta previsa vien più lenta."[1]

– Dante Alighieri, Divina Commedia, Paradiso

In Chapter 3, we introduced the novel conversational forecasting task of predicting whether a currently-civil conversation will derail into toxicity, and established its feasibility through a proof-of-concept baseline classifier that operates on linguistic features from the start of the conversation. In judging the overall utility of this baseline classifier, however, it is important to bear in mind our initial motivation (as laid out in Chapters 1 and 2): by forecasting the future derailment of a conversation based on early warning signs, we hope to give users and/or moderators enough *advance notice* to act before any harm is done (Liu et al., 2018; Zhang et al., 2018a; see Jurgens et al., 2019 for a discussion). In the context of this goal, however, the baseline classifier demonstrates clear shortcomings: it is restricted to a fixed set of features that may not generalize across domains, and more importantly, it extracts these features from only the first two comments of the conversation.

Overcoming these shortcomings is, however, nontrivial, as it requires addressing a unique property of conversations that make them distinct from other

[1] "The arrow one foresees arrives more gently."

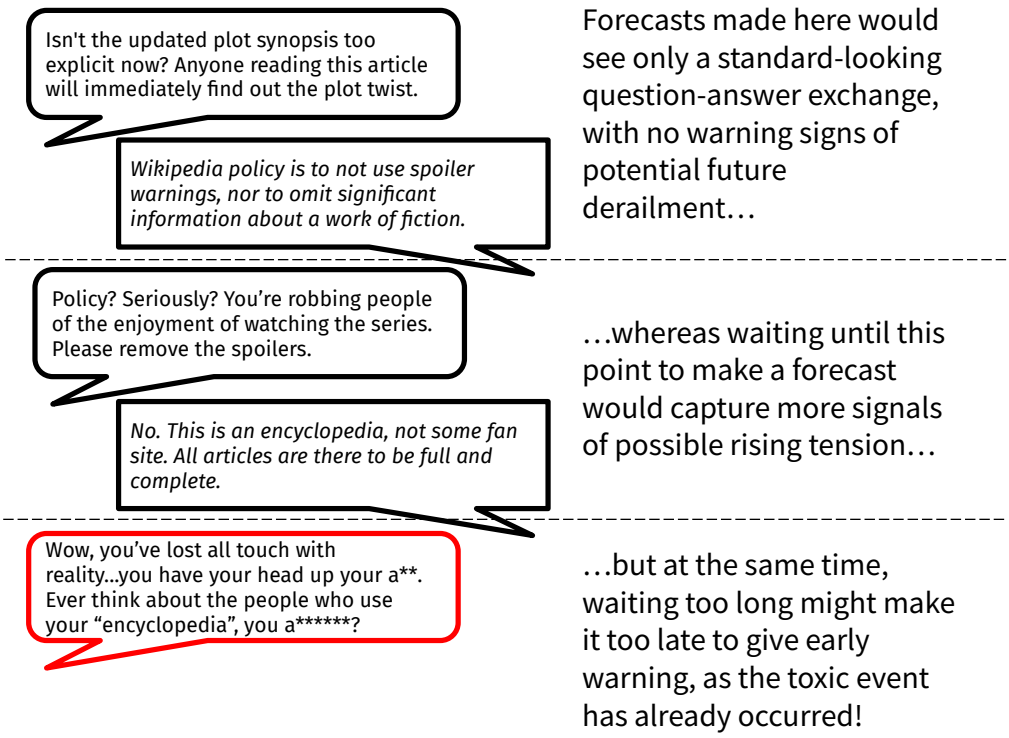


Figure 4.1: The inherent modeling challenges in practical forecasting of derailment, illustrated through an example conversation.

NLP domains: they evolve over time. This property leads to two inherent modeling challenges for forecasting (illustrated in Figure 4.1): it is impossible to know *in advance* (as would be required to provide users and moderators with *advance notice*) how many comments a conversation will get, and furthermore, each new comment has the potential to alter the trajectory of the conversation or recontextualize the meaning of comments that came before. In this chapter, we formalize these modeling challenges and argue that traditional NLP approaches—which are generally optimized for the case of static, unchanging documents—are ill-suited to address them.

Therefore, we argue, a brand-new class of models is needed—one that can overcome these inherent challenges by processing comments, and their rela-

tions, as they happen (i.e., in an online fashion), producing updated forecasts as the conversation evolves in real time. We accordingly refer to this new class of models as *conversational forecasting models*. Our main insight is that these core properties already exist in another class of models, albeit geared toward generation rather than forecasting: recent work in context-aware dialog generation (or "chatbots") has proposed sequential neural models that make effective use of the intra-conversational dynamics (Sordoni et al., 2015b; Serban et al., 2016, 2017), while concomitantly being able to process the conversation as it develops so that relevant responses can be generated even as the trajectory of the conversation changes (see Gao et al. (2018) for a survey).

In order for these systems to perform well in the generative domain they need to be trained on massive amounts of (unlabeled) conversational data. The main difficulty in directly adapting these models to the supervised domain of conversational forecasting is the relative scarcity of labeled data: for most forecasting tasks, at most a few thousands labeled examples are available, insufficient for the notoriously data-hungry sequential neural models.

To overcome this difficulty, we propose to decouple the objective of learning a neural representation of conversational dynamics from the objective of predicting future events. The former can be *pre-trained* on large amounts of unsupervised data, similarly to how chatbots are trained. The latter can piggyback on the resulting representation after *fine-tuning* it for classification using relatively small labeled data. While similar pre-train-then-fine-tune approaches have recently achieved state-of-the-art performance in a number of NLP tasks—including natural language inference, question answering, and commonsense reasoning (discussed in Section 4.3)—to the best of our knowledge this is the

first attempt at applying this paradigm to conversational forecasting. We implement this idea to create, to our knowledge, a first-of-its-kind conversational forecasting model, CRAFT (Section 4.5).

To test the effectiveness of this new architecture in forecasting derailment of online conversations, we develop and distribute two new datasets. The first triples in size the highly curated dataset that we introduced in Chapter 3, where civil-starting Wikipedia Talk Page conversations are crowd-labeled according to whether they eventually lead to personal attacks; the second relies on in-the-wild moderation of the popular subreddit ChangeMyView, where the aim is to forecast whether a discussion will later be subject to moderator action due to "rude or hostile" behavior.

Because conversational forecasting models follow conversations in real time and produce updated forecasts along the way, traditional machine learning metrics cannot be directly applied to evaluate them. Instead, we approach evaluation by first enumerating the desiderata of what constitutes a "good" forecast, then deriving from these a set of forecasting-specific definitions for the traditional notions of true (and false) positives (and negatives), from which it is possible to further derive adapted versions of familiar metrics like precision and recall. We evaluate CRAFT's performance on both datasets along these metrics. In both datasets, our model outperforms existing fixed-window approaches, as well as simpler sequential baselines that cannot account for inter-comment relations. Furthermore, by virtue of its online processing of the conversation, our system can provide substantial prior notice of upcoming derailment, triggering on average 3 comments (or 3 hours) before an overtly toxic comment is posted.

To summarize, in this chapter we:

- formalize the unique modeling challenges involved in practical forecasting of conversational events such as derailment;
- articulate the need for a new class of conversational forecasting models can capture the dynamics of a conversation in order to make updated forecasts *as the conversation develops*, and provide a concrete implementation of such a model;
- build two diverse datasets (one entirely new, one extending prior work) for the task of forecasting derailment of online conversations;
- design a framework for evaluating conversational forecasting models and use it to compare the performance of our model against baselines derived from prior work.

Our work is motivated by the goal of assisting users and moderators in online communities by preemptively signaling at-risk conversations that might require intervention to avoid derailment. However, we caution that any automated systems might encode or even amplify the biases existing in the training data (Park et al., 2018; Sap et al., 2019; Wiegand et al., 2019), so a public-facing implementation would need to be exhaustively scrutinized for such biases (Feldman et al., 2015).

Note on source material. This chapter is adapted from Chang and Danescu-Niculescu-Mizil (2019b). It expands on that work with more in-depth discussion on conversational forecasting and how it differs from traditional classification, additional baselines inspired by more recent work since the original paper was first published, and an analysis of variance in model performance for the sake of helping with reproducibility.

4.2 Conversational Forecasting

As defined in Chapter 3, conversational forecasting is a broad family of tasks with the shared goal of predicting future outcomes or events from a conversation; besides derailment which is our target of interest, other outcomes and events that have been studied include predicting the eventual success of persuasion attempts (Tan et al., 2016; Wachsmuth et al., 2018; Yang et al., 2019) and negotiations (Sokolova et al., 2008; Sicilia et al., 2024), the likelihood of future disagreements (Hessel and Lee, 2019), and the eventual decision that will be made after a team discussion (Smith, 2023). At first glance, forecasting may appear to just be a type of classification task, with the future event as a label; indeed (as we will discuss shortly) some early work on forecasting has operationalized it as classification. Yet we argue that conversational forecasting differs from traditional classification in two key ways, arising from a combination of the temporal aspect of forecasting and the difference between conversations and static documents. These differences, in turn, impose inherent yet often overlooked modeling challenges, requiring a new class of models to address them.

The first modeling challenge stems from the temporal dimension of conversational forecasting: that is, it aims to predict an event that has not yet occurred. The existence of this temporal dimension—standing in contrast to traditional classification, where the label is either known in advance or is an "atemporal" property of the document not associated with any specific point in time—raises a key question: when is a good time to make a forecast? The problem is that there is no universal answer to this question, because conversations have an *unknown horizon*: they can be of varying lengths, and the to-be-forecasted event can occur at any time. Because it is impossible to know in advance how long the

conversation will be and when (if at all) the event will occur, approaches that choose a fixed time to make a forecast (as our proof-of-concept from Chapter 3 did) may risk making a too-early prediction based on incomplete information or a too-late forecast that occurs after the event has already happened (and hence does not provide *early* warning), as illustrated in Figure 4.1. Subsequently, in order to truly generalize across all practical settings, a forecasting model cannot make a single forecast at a fixed point in time, but instead must follow a conversation in real time (i.e., in an online fashion) and make updated forecasts as the conversation evolves.

Compounding this is a second challenge arising from the fact that, in contrast to static documents, conversations are *dynamic*: the meaning of each new comment that comes in is not standalone, but rather may depend on the context of the preceding comments. Consequently, the overall trajectory of a conversation depends not only on the text of individual comments, but also on emergent properties that arise from the relationships between comments. Consider the behavior of the second user in Figure 4.1 (that is, the user whose contributions are marked with square text boxes). Intuitively, we might judge their behavior as somewhat stubborn and unwilling to compromise on their views regarding spoilers in the article—but notably, this fact cannot be inferred from the content of any *individual* comment of theirs, but instead arises from inter-comment patterns such as their use of a blunt "no" in response to the first user's request, and their repetitive citation of Wikipedia policy across all their comments. Thus, it is not merely the case that a forecasting model must follow the conversation in real-time—it must, on top of this, incorporate information from previous comments so that it can identify such inter-comment dynamics as they arise.

4.2.1 The Need for a New Class of Models

The two modeling challenges described above—that is, conversations being *dynamic* with *unknown horizon*—illustrate the differences between conversational forecasting and traditional classification. That having been said, prior work, including our own proof-of-concept in Chapter 3, has identified some compromising simplifications that can enable forecasting tasks to be reformulated as classification tasks, at the cost of practical applicability.

To address the challenge of conversational dynamics, a common simplification is to rely on hand-crafted features to capture such relations—e.g., similarity between comments (Althoff et al., 2016; Tan et al., 2016) or conversation structure (Zhang et al., 2018b; Hessel and Lee, 2019)—, though neural attention architectures have also recently shown promise (Jo et al., 2018). To address the challenge of unknown horizon, prior work has largely chosen from one of two possible simplifying assumptions. One approach is to assume (unrealistic) prior knowledge of when the to-be-forecasted event takes place and extract features up to that point (Niculae et al., 2015; Liu et al., 2018). An alternative is to extract features from a fixed-length window, often at the start of the conversation (Curhan and Pentland, 2007; Niculae and Danescu-Niculescu-Mizil, 2016; Althoff et al., 2016, inter alia), as we did in Chapter 3. Choosing a catch-all window-size is however inherently compromising: as we observed previously, short windows will miss information in comments they do not encompass, while longer windows risk missing the to-be-forecasted event altogether if it occurs before the end of the window. We note that such simplifications may sometimes take place more implicitly (and possibly unintentionally); for example, Kementchedjhieva and Sogaard (2021) implement a forecasting model

using BERT which, during preprocessing, truncates comments (to fit the BERT context window) recursively starting from the longest comment—which implicitly relies on prior knowledge of all utterances in the conversation so that the longest one can be identified.

Together, these compromising simplifications allow the forecasting task to be expressed as a classification task, which mathematically looks as follows:

$$p_{\text{event}} = f(\text{context}) \tag{4.1}$$

Where f is a standard classification algorithm (e.g., naive bayes, logistic regression, or more advanced neural models like multilayer perceptron), "context" is the hand-engineered representation of the conversation at a specific point in time that is fed as input to f, and p_{event} is the predicted probability of the to-be-forecasted event occurring, as output by f.

This approach certainly has value and should not be completely dismissed; indeed, it served as the basis of most of the work cited in this section as well as our own proof-of-concept in Chapter 3. Yet as we have argued, the compromises involved also prevent this approach from being used in a practical setting to provide early warning. For this latter use case, the key missing factor is the ability to adapt to changes in the conversation. To achieve this, we need to modify Equation 4.1 to be aware of the possibility of changes in the conversation—in other words, we need to introduce a temporal dimension:

$$p_{\text{event}}(t) = f(\text{context}(t)) \tag{4.2}$$

Here, the key change we have made is to turn both the input context and output p_{event} into functions of a timestamp t, which reflects the intuition that their value may change over time and factors in the key property that the to-be-forecasted event is in the future.

This abstract picture represents the novel class of model we ultimately want: a model that can perform the forecasting task in a real-time, online fashion, which we accordingly refer to as a *conversational forecasting model* (distinguishing it from traditional *classification models* that have been applied to forecasting tasks). The remainder of this chapter, then, is devoted to not only exploring how we can build such a model, but also how to evaluate it—which raises challenges of its own. In addition to presenting a concrete implementation of a conversational forecasting model and evaluating it, we position this novel class of models as a framework that other researchers and developers can use to address their own forecasting tasks, and offer via ConvoKit a general coding template that can be used to develop new conversational forecasting models.

4.3 Further Related Work

Antisocial behavior. Antisocial behavior online comes in many forms, including harassment (Vitak et al., 2017), cyberbullying (Singh et al., 2017), and general aggression (Kayany, 1998). Prior work has sought to understand different aspects of such behavior, including its effect on the communities where it happens (Collier and Bear, 2012; Arazy et al., 2013), the actors involved (Cheng et al., 2017; Volkova and Bell, 2017; Kumar et al., 2018; Ribeiro et al., 2018) and connections to the outside world (Olteanu et al., 2018).

Post-hoc classification of conversations. There is a rich body of prior work on classifying the outcome of a conversation after it has concluded, or classifying conversational events after they happened. Many examples exist, but some more closely related to our present work include identifying the winner

of a debate (Zhang et al., 2016; Potash and Rumshisky, 2017; Wang et al., 2017), identifying successful negotiations (Curhan and Pentland, 2007; Cadilhac et al., 2013), as well as detecting whether deception (Girlea et al., 2016; Pérez-Rosas et al., 2016; Levitan et al., 2018) or disagreement (Galley et al., 2004; Abbott et al., 2011; Allen et al., 2014; Wang and Cardie, 2014; Rosenthal and McKeown, 2015) has occurred.

Our goal is different because we wish to *forecast* conversational events before they happen and while the conversation is still ongoing (potentially allowing for interventions). Note that some post-hoc tasks can also be re-framed as forecasting tasks (assuming the existence of necessary labels); for instance, predicting whether an ongoing conversation *will* eventually spark disagreement (Hessel and Lee, 2019), rather than detecting already-existing disagreement.

Such hand-crafted features are typically extracted from fixed-length windows of the conversation, leaving unaddressed the problem of unknown horizon. While some work has trained *multiple* models for different window-lengths (Liu et al., 2018; Hessel and Lee, 2019), they consider these models to be independent and, as such, do not address the issue of aggregating them into a single forecast (i.e., deciding at what point to make a prediction). We implement a simple sliding windows solution as a baseline (Section 4.6).

Pre-training for NLP. The use of pre-training for natural language tasks has been growing in popularity after recent breakthroughs demonstrating improved performance on a wide array of benchmark tasks (Peters et al., 2018; Radford et al., 2018). Existing work has generally used a language modeling objective as the pre-training objective; examples include next-word prediction (Howard and Ruder, 2018), sentence autoencoding, (Dai and Le, 2015), and ma-

chine translation (McCann et al., 2017). More recent work has focused on greatly scaling up the language modeling that is done in pre-training, leading to a new family of models known as large language models (LLMs), which have demonstrated cutting-edge performance on a wide variety of NLP tasks (Bommasani et al., 2022; Wei et al., 2022). Within this new generation of models, the most similar in spirit to ours is BERT (Devlin et al., 2019), whose pre-training task is to predict the next sentence in a document given the current sentence. Our pre-training objective extends this concept to the *conversation* level (predicting the next utterance in a conversation) rather than a document level. We hence view our objective as *conversational modeling* rather than (only) language modeling.

4.4 Derailment Datasets

We consider two datasets, representing related but slightly different forecasting tasks. The first dataset is an expanded version of the annotated Wikipedia conversations dataset from Chapter 3. This dataset uses carefully-controlled crowdsourced labels, strictly filtered to ensure the conversations are civil up to the moment of a personal attack. This is a useful property for the purposes of model analysis, and hence we focus on this as our primary dataset. However, we are conscious of the possibility that these strict labels may not fully capture the kind of behavior that moderators care about in practice. We therefore introduce a secondary dataset, constructed from the subreddit ChangeMyView (CMV) that does not use post-hoc annotations. Instead, the prediction task is to forecast whether the conversation will be subject to moderator action in the future.

Wikipedia data (CGA-WIKI). In Chapter 3, we introduced a dataset of Wikipedia Talk Page conversations labeled by crowdworkers as either containing a personal attack from within (i.e., hostile behavior by one user in the conversation directed towards another) or remaining civil throughout. To the ends of more effective model training, we elected to expand this dataset, using the original annotation procedure. Since we found that the original data skewed towards shorter conversations, we focused this crowdsourcing run on longer conversations: ones with 4 or more comments preceding the attack.[2] Through this additional crowdsourcing, we expand the dataset to 4,188 conversations. We publicly release this data via ConvoKit (Chang et al., 2020b) as the "Conversations Gone Awry Dataset (Wikipedia Version)", or CGA-WIKI.

We perform a 60-20-20 train/dev/test split, ensuring that paired conversations end up in the same split in order to preserve the topic control. Finally, we randomly sample another 1 million talk page conversations to use for the unsupervised pre-training of the generative component.

Reddit CMV data (CGA-CMV). The CMV dataset is constructed from conversations collected via the Reddit API. In contrast to the Wikipedia-based dataset, we explicitly avoid the use of post-hoc annotation. Instead, we use as our label whether a conversation eventually had a comment removed by a moderator for violation of Rule 2: "Don't be rude or hostile to other users".[3]

Though the lack of post-hoc annotation limits the degree to which we can impose controls on the data (e.g., some conversations may contain toxic comments not flagged by the moderators) we do reproduce as many of the Wikipedia

[2]We cap the length at 10 to avoid overwhelming the crowdworkers.

[3]The existence of this specific rule, the standardized moderation messages and the civil character of the ChangeMyView subreddit was our initial motivation for choosing it.

data's controls as we can. Namely, we replicate the topic control pairing by choosing pairs of positive and negative examples that belong to the same top-level post, following Tan et al. (2016);[4] and enforce that the removed comment was made by a user who was previously involved in the conversation. This process results in 6,842 conversations, to which we again apply a pair-preserving 60-20-20 split. Finally, we gather over 600,000 conversations that do not include any removed comment, for unsupervised pre-training. As before, we publicly release this data via ConvoKit as the "Conversations Gone Awry Dataset (Reddit CMV version)", or CGA-CMV.

4.5 Online Forecasting Model

We now describe our concrete implementation of a conversational forecasting model. Our model integrates two components: (a) a generative dialog model that learns to represent conversational dynamics in an unsupervised fashion; and (b) a supervised component that fine-tunes this representation to forecast future events. Figure 4.2 provides an overview of the proposed architecture, henceforth CRAFT (Conversational Recurrent Architecture for ForecasTing).

Terminology. For modeling purposes, we treat a conversation as a sequence of N comments $C = \{c_1, \ldots, c_N\}$. Each comment, in turn, is a sequence of tokens, where the number of tokens may vary from comment to comment. For the n-th comment ($1 \leq n \leq N$), we let M_n denote the number of tokens. Then, a comment c_n can be represented as a sequence of M_n tokens: $c_n = \{w_1, \ldots, w_{M_n}\}$.[5]

[4] The top-level post is not part of the conversations.

[5] To keep model training computationally tractable, we cap the number of tokens at 80 (truncating anything beyond that) and the number of utterances at 16 (if a conversation ends up going longer than that, CRAFT discards comments in real time using a first-in-first-out scheme).

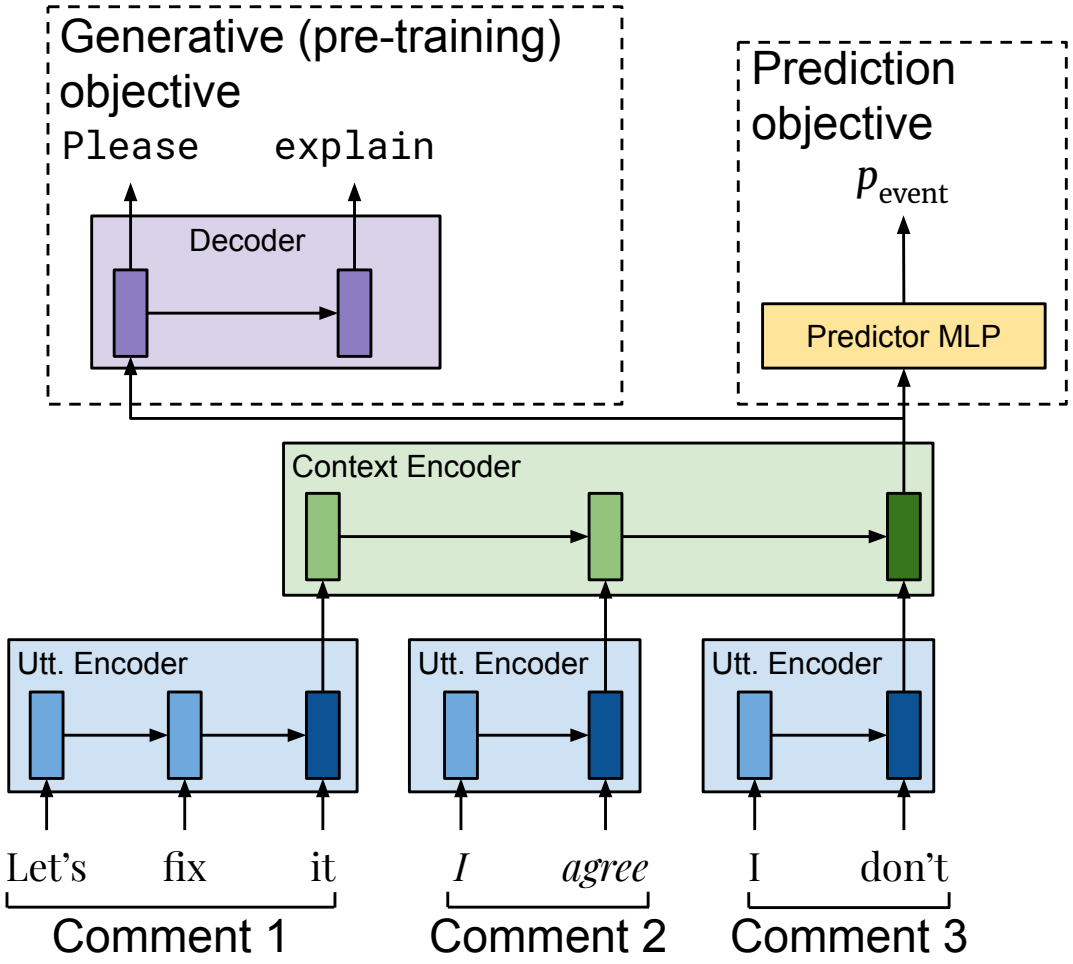


Figure 4.2: Sketch of the CRAFT architecture.

Generative component. For the generative component of our model, we use a hierarchical recurrent encoder-decoder (HRED) architecture (Sordoni et al., 2015a), a modified version of the popular sequence-to-sequence (seq2seq) architecture (Sutskever et al., 2014) designed to account for dependencies between consecutive inputs. Serban et al. (2016) showed that HRED can successfully model conversational context by encoding the temporal structure of previously seen comments, making it an ideal fit for our use case. Here, we provide a high-level summary of the HRED architecture, deferring deeper technical discussion to Sordoni et al. (2015a) and Serban et al. (2016).

An HRED dialog model consists of three components: an utterance encoder, a context encoder, and a decoder. The utterance encoder is responsible for generating semantic vector representations of comments. It consists of a recurrent neural network (RNN) that reads a comment token-by-token, and on each token w_m updates a hidden state h^{enc} based on the current token and the previous hidden state:

$$h_m^{\text{enc}} = f^{\text{RNN}}(h_{m-1}^{\text{enc}}, w_m) \tag{4.3}$$

where f^{RNN} is a nonlinear gating function (our implementation uses GRU (Cho et al., 2014)). The final hidden state h_M^{enc} can be viewed as a vector encoding of the entire comment.

Running the encoder on each comment c_n results in a sequence of N vector encodings. A second encoder, the context encoder, is then run over this sequence:

$$h_n^{\text{con}} = f^{\text{RNN}}(h_{n-1}^{\text{con}}, h_{M_n}^{\text{enc}}) \tag{4.4}$$

Each hidden state h_n^{con} can then be viewed as an encoding of the full conversational context up to and including the n-th comment. To generate a response to comment n, the context encoding h_n^{con} is used to initialize the hidden state h_0^{dec} of a decoder RNN. The decoder produces a response token by token using the following recurrence:

$$\begin{aligned} h_t^{\text{dec}} &= f^{\text{RNN}}(h_{t-1}^{\text{dec}}, w_{t-1}) \\ w_t &= f^{\text{out}}(h_t^{\text{dec}}) \end{aligned} \tag{4.5}$$

where f^{out} is some function that outputs a probability distribution over words; we implement this using a simple feedforward layer. In our implementation, we further augment the decoder with attention (Bahdanau et al., 2014; Luong et al., 2015) over context encoder states to help capture long-term inter-comment

dependencies. This generative component can be pre-trained using unlabeled conversational data.

Prediction component. Given a pre-trained HRED dialog model, we aim to extend the model to predict from the conversational context whether the to-be-forecasted event will occur. Our predictor consists of a multilayer perceptron (MLP) with 3 fully-connected layers, leaky ReLU activations between layers, and sigmoid activation for output. For each comment c_n, the predictor takes as input the context encoding h_n^{con} and forwards it through the MLP layers, resulting in an output score that is interpreted as a probability $p_{\text{event}}(c_{n+1})$ that the to-be-forecasted event will happen (e.g., that the conversation will derail):

$$p_{\text{event}}(c_n) = f^{\text{MLP}}(h_n^{\text{con}}) \tag{4.6}$$

We note that Equation 4.6 precisely mirrors the abstract picture of a conversational forecasting model defined in Equation 4.2, making it concrete by providing defined values as inputs: the multilayer perceptron f^{MLP} as the function f, and h_n^{con} as the context, to produce a forecast $p_{\text{event}}(c_n)$ that is specific to state of the conversation as of comment c_n (and therefore is time-dependent as specified by Equation 4.2).

Training the predictive component starts by initializing the weights of the encoders to the values learned in pre-training. The main training loop then works as follows: for each positive sample—i.e., a conversation containing an instance of the to-be-forecasted event (e.g., derailment) at comment c_e—we feed the context $c_1, \ldots, c_{e-1}$ through the encoder and classifier, and compute cross-entropy loss between the classifier output and expected output of 1. Similarly, for each negative sample—i.e., a conversation where none of the comments exhibit the to-be-forecasted event and that ends with c_N—we feed the context

$c_1, \ldots, c_{N-1}$ through the model and compute loss against an expected output of 0.

Note that the parameters of the generative component are not held fixed during this process; instead, backpropagation is allowed to go all the way through the encoder layers. This process, known as *fine-tuning*, reshapes the representation learned during pre-training to be more directly useful to prediction (Howard and Ruder, 2018).

We implement the model and training code using PyTorch. The full code is publicly available[6], and the trained models are also distributed via ConvoKit.

4.6 Forecasting Derailment

Having defined a concrete implementation of a conversational forecasting model, we now turn to the question of how well this model performs on our CGA-WIKI and CGA-CMV scenarios. Answering this question, however, requires addressing a more foundational question: how do we evaluate a forecaster? In other words, what makes a forecast "good"?

4.6.1 Defining Metrics for Evaluating Forecasts

As discussed in detail in Section 4.2, forecasting is different from traditional classification. The structural differences between them also mean that commonly-used metrics for evaluating classifiers are not directly applicable to evaluating

[6] `https://github.com/jpwchang/CRAFT`

conversational forecasting models. More concretely, recall the general mathematical representation of a classifier (Equation 4.1), wherein for a single (static) input document, the classifier produces one prediction. Evaluation typically proceeds by comparing this single prediction to a single ground-truth label, allowing the prediction to be categorized as, e.g., a true (or false) positive (or negative). From these categories, further metrics such as accuracy, precision, and recall can be derived.

By contrast, a conversational forecasting model produces *multiple* forecasts over the lifetime of a conversation. But each conversation still only has a single ground-truth label—in this case, whether or not it derails, but more generally, whether or not the to-be-forecasted event occurs—meaning we cannot simply treat each forecast as an independent prediction and apply standard classification metrics. Instead, we need to *aggregate* individual forecasts across the whole conversation to obtain a single point of comparison with the ground-truth label. But there are many possible ways to aggregate forecasts, ranging from basic approaches like taking a mean or median to more sophisticated combinations of multiple mathematical operations—which aggregation method is the right one to use? To answer this, we need to articulate exactly what it is we want; that is, to formally state the desiderata of what constitutes a "good" forecast.

To approach this, let us further reformulate the question: for any given conversation, what behavior would we want to see from an "ideal" forecaster? We consider two cases: conversations that actually derail, and ones that do not. If the conversation actually derails, we would intuitively like our forecasting model to, at *some* point during the conversation, forecast derailment. For now we will not be too picky about precisely when this forecast happens, since as

Figure 4.1 illustrates, signs of derailment may ebb and flow throughout the course of a conversation; instead we care only that at some point in the conversation, the model forecasts that the conversation will derail (though we will revisit this later in Section 4.7). Based on this intuition, we propose a forecasting-specific definition of **true positives (TP)** as actually-derailing conversations that are at some point forecasted to derail. Conversely, actually-derailing conversations that are never forecasted to derail are counted as **false negatives (FN)**. By similar reasoning, when considering a conversation that actually does not derail, the forecasting model should ideally never forecast that it will derail, and so such a case should be counted as a **true negative (TN)**.

What about the remaining situation, where a conversation does not derail but the forecasting model forecasts that it will derail? This case is somewhat trickier to reason about, because there is an argument to be made that a forecast of derailment in such a conversation is not necessarily "incorrect". Certainly, there are cases where the forecast was truly made in error and counts as a **false positive (FP)**, but on the other hand, it is also possible that there was at one point rising tension and conflict that legitimately put the conversation at risk of derailment, but this tension later somehow got defused. For example, imagine that after the fourth comment in Figure 4.1, a third party came in to mediate the emerging disagreement and successfully got the other users to calm down. This would change the subsequent trajectory of the conversation, but it would not retroactively change the fact that in the context of the first four comments there was rising tension, and a forecast of derailment made at that point is still arguably correct.[7] Given that there is no straightforward answer of how to ad-

[7] An analogy can be made to weather forecasting: if the forecast gives 70% chance of rain and it does not rain, this does not imply that the forecast was incorrect (as many a meteorologist will passionately remind you!).

dress this possibility, in this work we make the conservative assumption that all forecasts of derailment in non-derailing conversations are the result of model error and therefore count as false positives, which allows us to at least guarantee that we are not overestimating model performance. However, we emphasize that in practice, false positives are more complicated to reason about, and future work should examine the interesting yet complex phenomenon of conversations *recovering* from rising tension and conflict.

From these forecasting-specific definitions of true and false positives and negatives, it is possible to further derive the metrics of accuracy, precision, recall, F1, and false positive rate via their standard definitions. We evaluate the performance of CRAFT along these metrics on the CGA-WIKI and CGA-CMV datasets. To this end, for each of these datasets we pre-train the generative component on the corresponding unlabeled data and fine-tune it on the labeled training split (data size detailed in Section 4.4). These evaluation metrics are also built-in to ConvoKit's forecasting framework.

4.6.2 Baselines

Beyond just applying the metrics from Section 4.6.1 to CRAFT, we also wish to put these numbers in context by comparing to some baselines. Since CRAFT is a first-of-its-kind model, there is no directly comparable state-of-the-art to use as a natural baseline. Nonetheless, we can adapt techniques from standard NLP and from prior work (including Chapter 3) to create simpler conversational forecasting models to serve as reference points that CRAFT should ideally outperform.

Fixed-length window baselines. We first seek to compare CRAFT to existing, fixed-length window approaches to forecasting. To this end, we implement two such baselines: *Awry*, which is the proof-of-concept model introduced in Chapter 3, and *BoW*, a simple bag-of-words baseline that makes a prediction using TF-IDF weighted bag-of-words features extracted from the first comment-reply pair.

Online forecasting baselines. Next, we consider simpler approaches for making forecasts as the conversations happen (i.e., in an online fashion). First, we propose *Cumulative BoW*, a model that recomputes bag-of-words features on all comments seen thus far every time a new comment arrives. While this approach does exhibit the desired behavior of producing updated predictions for each new comment, it fails to account for relationships between comments.

This simple cumulative approach cannot be directly extended to models whose features are strictly based on a fixed number of comments, like Awry. An alternative is to use a *sliding window*: for a feature set based on a window of W comments, upon each new comment we can extract features from a window containing that comment and the $W-1$ comments preceding it. We apply this to the Awry method and call this model *Sliding Awry*. For both these baselines, we aggregate comment-level predictions in the same way as in our main model.

Off-the-shelf baselines. As mentioned in in Section 4.3, modern pretrained models such as BERT have shown competitive off-the-shelf performance on a wide range of NLP tasks. It is therefore natural to wonder whether they can be directly applied to forecasting derailment. To explore this, we implement a BERT baseline using hyperparameter values reported by Kementchedjhieva and Sogaard (2021). Our implementation differs from theirs in how we han-

dle BERT's fixed-size 512-token input window. Their implementation works by recursively stripping tokens from the longest comment, but as discussed in Section 4.2.1, this requires prior knowledge of what the longest comment is. To avoid assuming such prior knowledge, we instead adopt a sliding window approach like the one used for the Sliding Awry baseline, where after each new comment, the BERT baseline makes a forecast using as input the 512 most recent tokens in the concatenated comments of the conversation. We accordingly refer to this baseline as *Sliding BERT*.

CRAFT ablations. Finally, we consider two modified versions of the CRAFT model in order to evaluate the impact of two of its key components: (1) the pre-training step, and (2) its ability to capture inter-comment dependencies through its hierarchical memory.

To evaluate the impact of pre-training, we train the prediction component of CRAFT on only the labeled training data, without first pre-training the encoder layers with the unlabeled data. We find that given the relatively small size of labeled data, this baseline fails to successfully learn, and ends up performing at the level of random guessing.[8] This result underscores the need for the pre-training step that can make use of unlabeled data.

To evaluate the impact of the hierarchical memory, we implement a simplified version of CRAFT where the memory size of the context encoder is zero (*CRAFT − CE*), thus effectively acting as if the pre-training component is a vanilla seq2seq model. In other words, this model cannot capture inter-comment dependencies, and instead at each step makes a prediction based only on the utterance encoding of the latest comment.

[8]We thus exclude this baseline from the results summary.

Model	D	O	L	A	P	R	FPR	F1
BoW				56.5	55.6	65.5	52.4	60.1
Awry	✓			58.9	59.2	57.6	39.8	58.4
Cumul. BoW		✓		60.6	57.7	**79.3**	58.1	66.8
Sliding Awry	✓	✓		60.6	60.2	62.4	41.2	61.3
Sliding BERT	?	✓	✓	62.6	61.6	67.1	41.9	64.2
CRAFT – CE		✓	✓	64.9	**64.4**	66.7	**36.9**	65.5
CRAFT	✓	✓	✓	**66.5**	63.7	77.1	44.1	**69.8**

(a) CGA-WIKI

Model	D	O	L	A	P	R	FPR	F1
BoW				52.1	51.8	61.3	57.0	56.1
Awry	✓			54.4	55.0	48.3	**39.5**	51.4
Cumul. BoW		✓		59.9	58.8	65.9	46.2	62.1
Sliding Awry	✓	✓		56.8	56.6	58.2	44.6	57.4
Sliding BERT	?	✓	✓	**62.6**	**60.7**	71.8	46.5	65.8
CRAFT – CE		✓	✓	57.7	56.1	71.2	55.7	62.8
CRAFT	✓	✓	✓	62.1	59.0	**79.1**	55.0	**67.6**

(b) CGA-CMV

Table 4.1: Comparison of the capabilities of each baseline and our CRAFT models (full and without the Context Encoder) in both the (a) Wikipedia and (b) CMV settings. Models are compared in terms of their ability to capture inter-comment (D)ynamics, process conversations in an (O)nline fashion, and automatically (L)earn feature representations, as well as their performance in terms of (A)ccuracy, (P)recision, (R)ecall, False Positive Rate (FPR), and F1 score. 'Awry' is the baseline model from Chapter 3.

4.6.3 Results

Table 4.1 compares CRAFT to the baselines on the test splits (random baseline is 50%) and illustrates several key findings.[9] First, we find that unsurprisingly, accounting for full conversational context is indeed helpful, with even the

[9]Note that the (D)ynamics column is marked as '?' (for "unknown") on BERT; this is because while BERT does not *explicitly* model inter-comment dynamics like CRAFT does, one may speculate that it could still *implicitly* learn such dynamics by seeing conversational data in its pre-training. The black-box nature of this model leaves this speculation inconclusive.

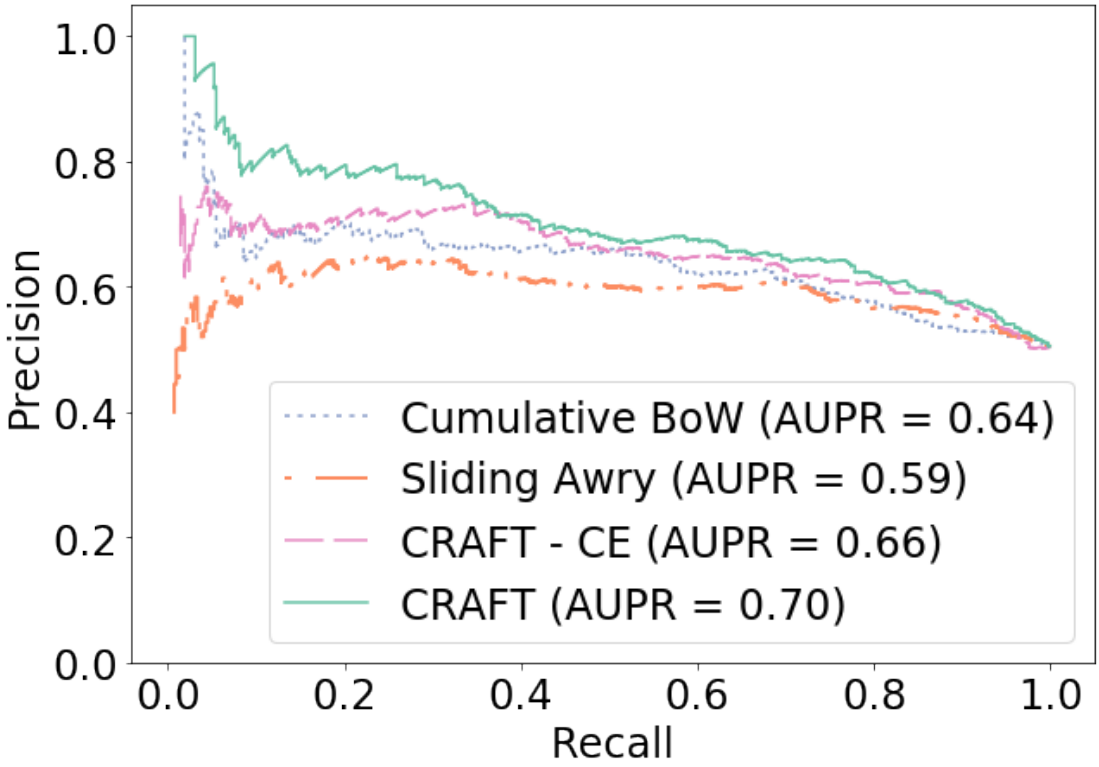
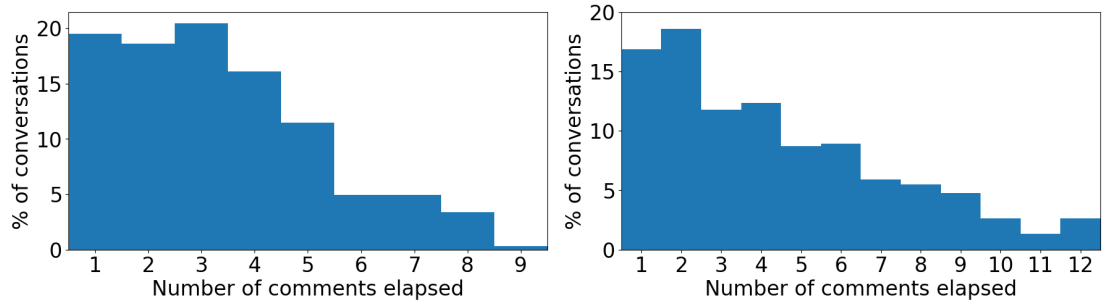


Figure 4.3: Precision-recall curves and the area under each curve. To reduce clutter, we show only the curves for Wikipedia data (CMV curves are similar) and exclude the fixed-length window baselines (which perform worse).

simple online baselines outperforming the fixed-window baselines. On both datasets, CRAFT outperforms all the non-BERT baselines (including the other online models) in terms of accuracy and F1. Furthermore, although it loses on precision (to CRAFT − CE) and recall (to Cumulative BoW) individually on CGA-WIKI, CRAFT has the superior *balance* between the two, having both a visibly higher precision-recall curve and larger area under the curve (AUPR) than the baselines (Figure 4.3). This latter property is particularly useful in a practical setting, as it allows moderators to tune model performance to some desired precision without having to sacrifice as much in the way of recall (or vice versa) compared to the baselines and pre-existing solutions.

In contrast to the other baselines, Sliding BERT is—at least in the CGA-CMV setting—more closely competitive with CRAFT, with the two models being roughly on par in accuracy and precision. We note, however, that CRAFT does noticeably higher on recall (in both datasets) without having to trade off precision. This property is again potentially useful in the context of balancing precision and recall in a practical setting. Overall, we argue that this shows that despite the generally good off-the-shelf performance of pretrained models on most NLP tasks, the inherent modeling challenges involved in forecasting (Section 4.2) pose a challenge even to these state-of-the-art general-purpose models, and there is still value in building models like CRAFT that are specifically geared towards forecasting.

In interpreting these results, it is important to bear in mind that, as neural models, CRAFT and BERT performance may vary due to nondeterminism in the training process. To quanitfy this variance, we perform 10 runs of CRAFT and BERT fine-tuning from scratch on each dataset, from which we can compute the standard deviation of each metric. We find that CRAFT's overall variance is relatively low, with accuracy and F1 both having standard deviations of $< 1\%$, meaning those results are robust. That said, it should be noted that precision, recall, and false positive rate are somewhat more variable: they have standard deviations between $1 - 3\%$ on CGA-WIKI, or $1 - 5\%$ on CGA-CMV (the latter finding being perhaps a consequence of the relative noisiness of the CGA-CMV data). By contrast, BERT's performance is somewhat more variable, with $> 1\%$ standard deviation on F1. Despite the variance in both models, however, our earlier comparisons generally still hold across runs. The full results of all 10 runs on each dataset can be found in Appendix E.



(a) (b)

Figure 4.4: Distribution of number of comments elapsed between the model's first warning and the toxic comment in the (a) CGA-WIKI and (b) CGA-CMV scenarios.

4.7 Analysis

We now examine the behavior of CRAFT in greater detail, to better understand its benefits and limitations. We specifically address the following questions: (1) How much early warning does the the model provide? (2) How does conversation length affect model performance? (3) Does the model actually learn an order-sensitive representation of conversational context?[10]

Early warning, but how early? The recent interest in forecasting derailment has been driven by a desire to provide pre-emptive, actionable warning to moderators. But does our model trigger early enough for any such practical goals?

For each actually-derailing conversation that the model correctly forecasts will derail (i.e., for each true positive), we count the number of comments elapsed between the time the model is first triggered and the actual toxic comment; note that by definition this metric is *only* computable on true positives, as it requires the existence of both an actual toxic comment and a trigger (i.e. a pos-

[10]We choose to focus on CGA-WIKI since the conversational prefixes are hand-verified to be civil. For completeness we also report results for CGA-CMV throughout, but they should be taken with an additional grain of salt.

Subset	A	P	R	FPR	F1
1st quartile (< 5 comments)	71.4	70.6	73.3	30.5	72.0
2nd quartile (5 – 6 comments)	67.7	64.4	79.2	43.8	71.0
3rd quartile (6 – 8 comments)	63.8	60.5	79.7	52.0	68.8
4th quartile (≥ 8 comments)	60.5	58.0	75.8	54.8	65.7

(a) CGA-WIKI

Subset	A	P	R	FPR	F1
1st quartile (< 4 comments)	67.1	66.7	68.4	34.2	67.5
2nd quartile (4 – 6 comments)	64.6	61.9	75.8	46.7	68.1
3rd quartile (6 – 8 comments)	59.9	56.3	87.6	67.9	68.6
4th quartile (≥ 8 comments)	55.9	53.6	87.1	75.3	66.4

(b) CGA-CMV

Table 4.2: Performance of CRAFT on subsets of the (a) CGA-WIKI and (b) CGA-CMV test sets, subdivided by conversation length.

itive forecast). Figure 4.4 shows the distribution of these counts: on average, the model warns of toxicity 3 comments before it actually happens (4 comments for CMV). To further evaluate how much time this early warning would give to the moderator, we also consider the difference in timestamps between the comment where the model first triggers and the actual toxic comment. Over 50% of conversations get at least 3 hours of advance warning (2 hours for CMV). Moreover, 39% of conversations get at least *12 hours* of early warning before they derail.

How does performance vary with conversation length? The above analysis sheds light on how the exact timing of CRAFT's initial forecast can vary, but it is (by design) independent of the actual length of the conversation. A natural follow-up question, then, is whether the length of a conversation matters in terms of how well CRAFT does at forecasting. Notably, if length does turn out to affect CRAFT's performance, it is not immediately obvious a priori what direction we should expect the effect to go in. On the one hand, it is plausible that longer conversations offer more information for CRAFT to base its decisions on,

and would thereby lead to increased performance. On the other hand, it is also plausible that longer conversations offer more opportunities for CRAFT to get things wrong, and would thereby lead to decreased performance.

To settle this question, we take the CGA-WIKI test set and subdivide it based on length. The subdivision is done by quartiles, so that we end up with four approximately evenly-sized subsets: conversations with fewer than 5 comments in the first quartile, 5-6 comments in the second quartile, 6-8 comments in the third quartile, and 8 or more comments in the fourth quartile. We then compute the performance metrics within each subset. We also repeat this experiment on the CGA-CMV test set, which has slightly different (but broadly similar) cutoff points for its quartiles.

The results of this analysis are shown in Figure 4.2. In both settings, we find that there is indeed a length effect, and that specifically CRAFT has better performance on shorter conversations. As conversation length increases, we observe decreases in accuracy, precision, and F1, and increases in false positive rate (more pronounced for CGA-CMV than for CGA-WIKI). These findings may have implications for real-world usage of CRAFT; for instance, given the reduced performance on longer conversations, applications using CRAFT may consider imposing a higher decision threshold once the conversation passes a certain length. Future work should explore such possibilities and empirically evaluate what settings offer the best overall forecasting power.

Does order matter? One motivation behind the design of our model was the intuition that comments in a conversation are not independent events; rather, the order in which they appear matters (e.g., a blunt comment followed by a polite one feels intuitively different from a polite comment followed by a blunt one).

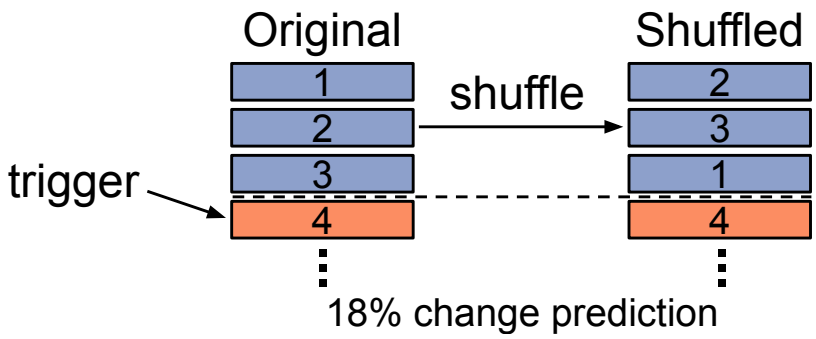


Figure 4.5: The prefix-shuffling procedure ($t = 4$).

By design, CRAFT has the capacity to learn an order-sensitive representation of conversational context, but how can we know that this capacity is actually used? It is conceivable that the model is simply computing an order-insensitive "bag-of-features". Neural network models are notorious for their lack of transparency, precluding an analysis of how *exactly* CRAFT models conversational context. Nevertheless, through two simple exploratory experiments, we seek to show that it does not completely ignore comment order.

The first experiment for testing whether the model accounts for comment order is a *prefix-shuffling* experiment, visualized in Figure 4.5. For each conversation that the model predicts will derail, let t denote the index of the triggering comment, i.e., the index where the model first made a derailment forecast. We then construct *synthetic* conversations by taking the first $t - 1$ comments (henceforth referred to as the *prefix*) and randomizing their order.[11] Finally, we count how often the model no longer predicts derailment at index t in the synthetic conversations. If the model were ignoring comment order, its prediction should remain unchanged (as it remains for the Cumulative BoW baseline), since the actual *content* of the first t comments has not changed (and CRAFT inference is deterministic). We instead find that in roughly one fifth of cases (12% for

[11] We restrict the experiment to cases where $t \geq 3$, as prefixes consisting of only one comment cannot be reordered.

CMV) the model changes its prediction on the synthetic conversations. This suggests that CRAFT learns an order-sensitive representation of context, not a mere "bag-of-features".

To more concretely quantify how much this order-sensitive context modeling helps with prediction, we can actively prevent the model from learning and exploiting any order-related dynamics. We achieve this through another type of shuffling experiment, where we go back even further and shuffle the comment order in the conversations used for pre-training, fine-tuning and testing. This procedure preserves the model's ability to capture signals present within the individual comments processed so far, as the utterance encoder is unaffected, but inhibits it from capturing any meaningful order-sensitive dynamics. We find that this hurts the model's performance (65% accuracy for Wikipedia, 59.5% for CMV), lowering it to a level similar to that of the version where we completely disable the context encoder.

Taken together, these experiments provide evidence that CRAFT uses its capacity to model conversational context in an order-sensitive fashion, and that it makes effective use of the dynamics within. An important avenue for future work would be developing more transparent models that can shed light on exactly *what* features are being extracted and *how* they are used in prediction.

4.8 Conclusions and Discussion

In this chapter, we articulated a need for a new class of *conversational forecasting models* that can process comments as they happen and take the full conversational context into account to make an updated forecast at each step. We then

described CRAFT, a first-of-its-kind concrete implementation of this idea. This model fills a void in the existing literature on conversational forecasting, simultaneously addressing the dual challenges of capturing inter-comment dynamics and dealing with an unknown horizon. We find that our model achieves not only beats naive baselines on the task of forecasting derailment in two different datasets that we release publicly, but even goes neck-and-neck with—or in some cases exceeds—the performance of the popular pretrained model BERT. We further show that the resulting system can provide substantial prior notice of derailment, opening up the potential to assist in proactive interventions to keep the conversation on track (Seering et al., 2017).

While we have focused specifically on the task of forecasting derailment, we view this work as a step towards a more general model for real-time forecasting of other types of emergent properties of conversations. Follow-up work could adapt the CRAFT architecture to address other forecasting tasks mentioned in Section 4.3—including those for which the outcome is extraneous to the conversation. We expect different tasks to be informed by different types of inter-comment dynamics, and further architecture extensions could add additional supervised fine-tuning in order to direct it to focus on specific dynamics that might be relevant to the task (e.g., exchange of ideas between interlocutors or stonewalling). To support such efforts, we have implemented a general framework for developing conversational forecasting models as part of ConvoKit, enabling the standardization of future models such that they can easily be interchanged and directly compared.

A practical limitation of the current analysis is that it relies on balanced datasets, while derailment is a relatively rare event for which a more restrictive

trigger threshold would be appropriate. While our analysis of the precision-recall curve suggests the system is robust across multiple thresholds ($AUPR = 0.7$), additional work is needed to establish whether the recall tradeoff would be acceptable in practice.

Additionally, one major limitation of the present work is that it assigns a single label to each conversation: does it derail or not? In reality, derailment need not spell the end of a conversation; it is possible that a conversation could get back on track, suffer a repeat occurrence of antisocial behavior, or any number of other trajectories. It would be exciting to consider finer-grained forecasting of conversational trajectories, accounting for the natural—and sometimes chaotic—ebb-and-flow of human interactions.

Finally, there remain open questions regarding what human users and moderators actually desire from an early-warning system, which would affect the design of a practical system based on this work. For instance, how early does a warning need to be in order for users or moderators to find it useful? What is the optimal balance between precision, recall, and false positive rate at which such a system is truly helping humans find derailing conversations they might have otherwise missed, while also avoiding wasting their time through false positives? What are the ethical implications of such a system? We regard these questions as an important part of the overall evaluation of any derailment forecasting model, since assisting users and moderators was the original underlying motivation. Answering these questions requires integrating CRAFT into a prototype tool that is accessible to laypersons, and conducting user studies to understand how users and moderators interact with such a tool—a nontrivial undertaking that constitutes the focus of the next chapter.

CHAPTER 5

HOW FORECASTING DERAILMENT CAN HELP ONLINE COMMUNITIES

5.1 Introduction

Our key findings up to now have been as follows:

- In Chapter 2, we found that users and moderators of online communities have some intuition for when conversations might be at risk of derailing into toxicity, and also have some strategies for proactively preventing such an outcome. At the same time, we also found that this intuition is imperfect and that users and moderators alike express uncertainty about whether and when it is appropriate to act proactively.
- In Chapter 3, we formalized the task of identifying conversations that are at risk of derailment as a novel conversational forecasting task and showed that this task is feasible for computational methods. We followed this up in Chapter 4 with a new model for practical, real-time forecasting of derailment.

In other words, we have identified both a **problem**—users and moderators face practical challenges in knowing when to take proactive action to prevent a conversation from derailing into toxicity—and a potential **solution**—computational approaches, which are capable of capturing some of the human intuition about risk of derailment, could provide forecasts to help augment existing human judgment; in other words, enhance users' *risk awareness*.

Yet thus far, this potential solution has been purely hypothetical. While we have found that our CRAFT model performs well along a series of metrics for evaluating forecasts (Section 4.6), it is well-documented that these metrics do not perfectly correlate with real-world utility (Ribeiro et al., 2020; Raji et al., 2022). Therefore, the aim of this chapter is to go beyond just evaluating the CRAFT model itself, and instead evaluate its utility in our proposed proactive risk awareness paradigm: can forecasts from CRAFT effectively improve users' experiences in their online discussions? To narrow the scope of our investigation, we will focus for now on the perspective of ordinary users, though we will also provide concrete pointers for how this work could be extended to apply to moderators as well (Section 5.5).

In designing this evaluation, a top priority for us is to gain as realistic as possible a picture of how real users react to algorithmic forecasts in a real setting; that is, in the standard parlance of study design, we aim to prioritize ecological validity and avoid the known distorting effects of laboratory or crowdsourced studies (Taylor et al., 2019; Reinecke and Gajos, 2015; Goodman et al., 2013; Nichols and Maner, 2008). Accordingly, we follow the precedent of similar work in human-computer interaction (Katsaros et al., 2022; Kohlbrenner et al., 2022) and conduct an "in-the-wild" user study in a popular online discussion platform, ChangeMyView. Executing this requires us to engage with several core challenges: not only technical, but also practical and ethical.

From a technical perspective, we need to build a system that can apply CRAFT (or any other forecasting algorithm) to real ongoing discussions, and generate forecasts in real time so that it can inform users about the risk of derailment and about the potential impact of their responses *as they are drafting them*.

To this end we develop a prototype tool, which we call ConvoWizard, consisting of two interoperating components: a *backend* algorithmic scoring system powered by CRAFT, and a *frontend* browser plugin that relays conversational data to the backend and, based on the results returned by the CRAFT backend, serves interventions to users directly on the ChangeMyView webpage.[1]

From an ethical and practical perspective, turning regular platform users into volunteers for a scientific study requires a design that puts their needs and well-being at the fore. Thus, in designing and conducting this user study, we adopted a community collaboration model which took direct input from ChangeMyView community leaders. Additionally, we used a two-phase study design, starting with a larger phase in which we sought feedback from the participants after using the fully functional tool for one month, and continuing with a second phase in which we implemented a within-participant randomized controlled experiment lasting two months.

The results of the user study suggest that the risk awareness paradigm has the potential to improve online discourse and motivate further research in this direction. In exit surveys, the majority of participants reportthat they found ConvoWizard helpful for identifying tense situations, with the tool both *supplementing* their intuitions—catching types of tension that they may not have known to look for—and *activating* their existing intuitions—reminding them to be on the lookout for tension in situations where they may not have been paying attention. Most participants also report that this additional awareness of risk helped them avoid fights and kept them from posting comments they would have regretted later.

[1] A video demostration of ConvoWizard is available at `https://www.cs.cornell.edu/~cristian/Thread_With_Caution.html`.

Combining feedback from participants with quantitative analysis of the data from the randomized controlled experiment offers a glimpse into concrete steps participants took in response to this increased awareness. First, participants report that seeing a warning from ConvoWizard led them to reflect more on the tension in the conversation and how their reply might affect it. This effect is echoed in the randomized controlled experiment results: when users are warned that a conversation they are participating in is at risk of future incivility, they spend 9% more time on average drafting their comment compared to the control condition where they are not warned of this risk. Beyond reflecting on tension, participants further report that they go on to revise their draft reply using ConvoWizard as a guide to reduce the risk of derailment. This effect is again echoed in the experimental results: when users are warned about an existing risk they edit their reply in a way that tends to gradually decrease this risk, whereas in the control condition where they are not warned, they tend to escalate the risk. While the observed effects are small and limited by the scale of our study, they nonetheless combine with our qualitative observations to offer a promising initial indicator that directly empowering well-intentioned users with additional awareness about risk of derailment is a feasible complement to existing moderation practices, with potential to improve online discourse. This establishes a groundwork for future studies by highlighting concrete directions for future implementations of this paradigm.

In summary, in this chapter we:

- propose a new paradigm that empowers well-intentioned users to assess and address the risk of incivility in the conversations they participate in;
- develop a fully functional tool that implements this paradigm in a popular

discussion community; and

- design and conduct a user study, in collaboration with the moderators of this community, to evaluate the feasibility and potential of this new paradigm.

Note on source material. This chapter adapts and synthesizes material from Schluger et al. (2022) and Chang et al. (2022).

5.2 Related Work

5.2.1 User-facing Interventions

Our current work fits into the broader landscape of *user-facing interventions*: meant to steer users towards more pro-social behaviors. Such technology is built upon the principle that ordinary users can and should play a role—alongside traditional moderation—in community governance and norm maintenance (as we have previously outlined in Chapter 1).

The concept of using user-facing interventions to guide and promote pro-social behavior descends from earlier work in HCI which has studied how interventions might be used to promote *offline* behaviors, most notably in the area of health and fitness (Krebs et al., 2010). Among the earliest work on interventions directly related to building and maintaining pro-social norms is Kriplean et al. (2012a)'s "ConsiderIt", which presents an experimental platform for political deliberation in which users are guided to critically examine talking points from both sides and attempt to better understand each others' perspectives. Con-

siderIt could be viewed as an early attempt at using platform design to reduce the likelihood of misunderstandings (a potential factor in derailment, as we describe in Section 2.2); this goal is more explicitly tackled in its contemporary work "Reflect" (Kriplean et al., 2012b), a proof-of-concept discussion platform where users seeking to reply to a comment are asked to first reflect on what the commenter might have meant and to explain it in their own words.

While the interventions tested in ConsiderIt and React involved prominently asking users to go out of their way to complete intensive tasks, more recent work has explored more subtle approaches. Seering et al. (2019a) build upon the foundation laid by ConsiderIt and React, but reformulate the intervention as simpler CAPTCHA-style tasks meant to promote pro-social behavior on a more subconscious level. Taylor et al. (2019) take this subconscious approach even further, introducing "empathy nudges" in the form of minor UI adjustments—for instance, personalizing the reply button by showing the name of the user being replied to—that do not explicitly require any extra action from the user.

While experiments with these interventions have shown success, the authors of one such system, Taylor et al. (2019), caution that their implementation (and others like it) suffer from a key vulnerability that might limit their effectiveness in the real world: they are *static*, in the sense that they are globally applied across all of a user's interactions. The problem is that not every interaction requires an intervention; in most interactions people are already behaving civilly. Though at first this seems at most a minor annoyance, Taylor et al. reason that because that peoples' capacity for empathy is finite, static interventions might only work in a limited lab setting—if users were seeing the intervention all the time in their everyday social media usage, they might get overwhelmed and

just tune it out. A similar "attrition" effect, where static interventions lose effectiveness over time when deployed at scale, has been observed in work on interventions in other fields (Krebs et al., 2010; Kovacs et al., 2018; Collins et al., 2014). Thus, Taylor et al. argue, making proactive interventions effective at scale requires a *dynamic* approach of "targeting design interventions just in time for the individuals who need them."

Up until recently, technical barriers have stood in the way of developing such dynamic interventions: in lieu of a method to automatically identify situations where interventions are needed, what little work has been done in this direction in the past has relied upon simple heuristics to decide when interventions are shown. For instance, Halfaker et al. (2011b) test an intervention on Wikipedia that is only shown in interactions involving a user who is new to the community, under the assumption that these are particularly sensitive situations where toxicity could impact the new user's likelihood of remaining active on Wikipedia. But such simple heuristics may not capture all situations where interventions are needed—and with toxicity only continuing to grow as a problem in online communities, this has led to recent calls for the NLP community to investigate more sophisticated approaches for automatically detecting, and intervening in, conversations at risk of turning toxic (Jurgens et al., 2019).

Our present work adds to the ongoing research on proactive intervention design by exploring an initial response to Taylor et al. (2019)'s and Jurgens et al. (2019)'s calls for targeted, just-in-time interventions. We believe that derailment forecasting algorithms, like the CRAFT model we introduced in Chapter 4, can fill the technical gap to make such dynamic interventions possible, automatically identifying interactions in need of intervention by detecting *rising tension*

that could lead to future toxicity. We use this technology to build our own implementation of a dynamic, proactive intervention system (Section 5.3.1).

5.2.2 "In-the-wild" Study Design

Our approach to evaluating our proposed proactive intervention system similarly derives from lessons learned in prior work on interventions and HCI more broadly. By far the most common approach to evaluating interventions has been the use of laboratory studies, in which participants—recruited either by traditional means or by crowdsourcing—are asked to specifically try out an intervention, which may be implemented as either a simulation/mockup, or as a fully implemented app that is deployed only within the confines of the experiment (i.e., not open to the public). Variations of this approach were used in most of the previously described intervention studies: Seering et al. (2017) and Taylor et al. (2019) recruited crowdworkers to evaluate mocked-up designs of their interventions, while Kriplean et al. (2012a)'s "ConsiderIt" was a fully-implemented system tested in an isolated in-person setting.

However, the artificial nature of laboratory settings has raised questions over the extent to which laboratory-observed effects translate to real-world effects. Research within both HCI and psychology has uncovered various drawbacks of laboratory studies that might limit the ecological validity of their findings: experimental subjects are often prone to biases such as tending to give responses that they think the experimenter wants (Nichols and Maner, 2008; Klein et al., 2012), and in the specific case of studying interventions, some interventions might be effective during the limited time window of the experiment—when

users are paying undivided attention to the intervention—but lose effectiveness in more realistic scenarios where users are juggling other interests, concerns, and distractions (Collins et al., 2014; Kovacs et al., 2018). To avoid such effects an gain a truer picture of an intervention's effects, it is necessary to go outside the laboratory and run studies in real-world settings that more accurately reflect how the user would actually interact with the intervention on a day-to-day basis—in other words, to run studies "in the wild" (Consolvo et al., 2008; Reinecke and Gajos, 2015; Mottelson and Hornbæk, 2017).

Running studies "in the wild" also brings its own set of challenges, however. The bulk of online interactions today take place on closed-source, proprietary platforms that reveal little about their inner workings (Paterson, 2012), making it difficult or downright impossible to study user behaviors on those platforms. While these platforms sometimes engage in collaborations with the research community to make experimentation possible, they generally impose limitations on these collaborations that limit the scope of what can be done and raise questions about the validity of the resulting findings (Morstatter et al., 2013; Allen et al., 2021). One way to circumvent the limitations of major proprietary platforms is for researchers to develop their own custom platform which they have full access to and control over, as Kriplean et al. (2012b) did with their "Reflect" discussion platform. However, recruiting users to seriously use such a platform (and not just treat it as an experimental setting) remains challenging, largely limiting the scope of this approach to specific niches such as online games (Fu et al., 2017).

Recently, there has been increasing attention towards an alternative solution for in-the-wild studies which enable researchers to tap into major online plat-

forms while avoiding some of the limitations associated with those platforms: collecting data via *browser extensions* (Kohlbrenner et al., 2022; Ali et al., 2023). This approach aims to offer the best of both worlds: researchers maintain full control over browser extensions that they develop, much like they would with a custom platform, while at the same time extensions can integrate tightly with the browser in order to provide a window into otherwise-inaccessible user behaviors on closed platforms. Furthermore, the opt-in nature of the extension approach helps to avoid concerns about informed consent that can arise from platform-run experiments (Jouhki et al., 2016). Our present work adopts this approach, implementing an intervention for our risk awareness paradigm that is served to users through a custom browser extension, whose design we describe in Section 5.3.1.

5.3 Methods

To evaluate the feasibility of our proposed risk awareness paradigm we develop a prototype tool that implements it, ConvoWizard (Section 5.3.1), and gather both qualitative feedback and quantitative usage data through an IRB-approved user study (Section 5.3.2). Following a rich line of work on in-the-wild study design (Section 5.2.2), we set up our user study to involve real users in an actual social media community, namely the Reddit debate forum ChangeMyView (previously discussed in Section 2.3.1). However, the real-world setting also introduces a host of technical, practical, and ethical challenges, which end up shaping the design of our study:

Technical challenge: How can we provide users with real-time information

about the risk of real online conversations? In a laboratory setting, the researchers would have full control over both the conversations that get shown (which would enable them to pre-annotate the risk of each conversation) and the UI of the simulated platform (which would enable them to easily add the risk information as an additional UI element). By contrast, real online conversations take place on established platforms that we lack control over. In Section 5.3.1, we explain the technical approach we take to tackling this problem: developing a browser extension that reads the content of conversations taking place on Reddit, uses CRAFT to algorithmically score the risk level of that content in real time, and extends the Reddit UI with additional elements that can be used to display interventions based on the score.

Practical challenge: How can we convince everyday users of online platforms to use our tool as part of their regular activity? In particular, since we implement our interventions via a browser extension, participants need to be willing to not only install the software but also keep it enabled for the full duration of the study. Therefore, the tool needs to provide real value to the user in addition to supporting the research. In section 5.3.2, we explain how we set up the experimental conditions in order to combine these goals.

Ethical challenge: Algorithmic systems can produce flawed or biased judgments (Davidson et al., 2017; Duarte and Llansó, 2018), and harm can occur if such flawed judgments are used as the basis of real-world actions. In the specific context of our study, this could take the form of our tool providing wrong estimates of risk to users, which might cause them to make bad decisions. Because our study is taking place in real online discussions, the potential harm is not just limited to the study participants themselves, but to other users in the

discussion, and perhaps even the broader community. This danger carries a clear ethical implication: because the community shoulders the potential harms arising from flaws or misuse of our technology, the community should be consulted and involved in the running of the study. This conclusion leads us to develop our study as a *community collaboration*, done as a joint endeavor with the moderators of ChangeMyView. In Section 5.3.2, we explain this approach in more detail.

5.3.1 Technical Design: The ConvoWizard Tool

To address the technical challenge of presenting users with advance notification of how their comments may affect a conversation, we build ConvoWizard: a prototype tool that is designed to assess the risk of conversations in real time and deliver this information to the user. ConvoWizard is comprised of two parts: (1) a browser extension we distributed to participants in the study which extracts data about the conversations they engage with on ChangeMyView, collects data about their in-progress drafts, and displays UI interventions; and (2) a backend server which runs CRAFT in real time to predict the trajectory of ongoing conversations, relays this information to the browser extension, and logs data for subsequent analysis.

Frontend: the user facing extension

ConvoWizard's user-facing frontend is implemented as a Google Chrome extension which operates by reading and manipulating Reddit's browser-side HTML

Figure 5.1: The Context Summary feature of ConvoWizard provides information about whether the conversation the user is joining is at risk of turning uncivil in the future. (b) When no risk is detected, the Context Summary displays a neutral message on a blank background. (c) When risk is detected, the Context Summary displays a warning message displayed on a red background, with deeper shades of red indicating higher risk. Note that both examples come from the same discussion thread; for reference, the post that started the thread is shown in (a).

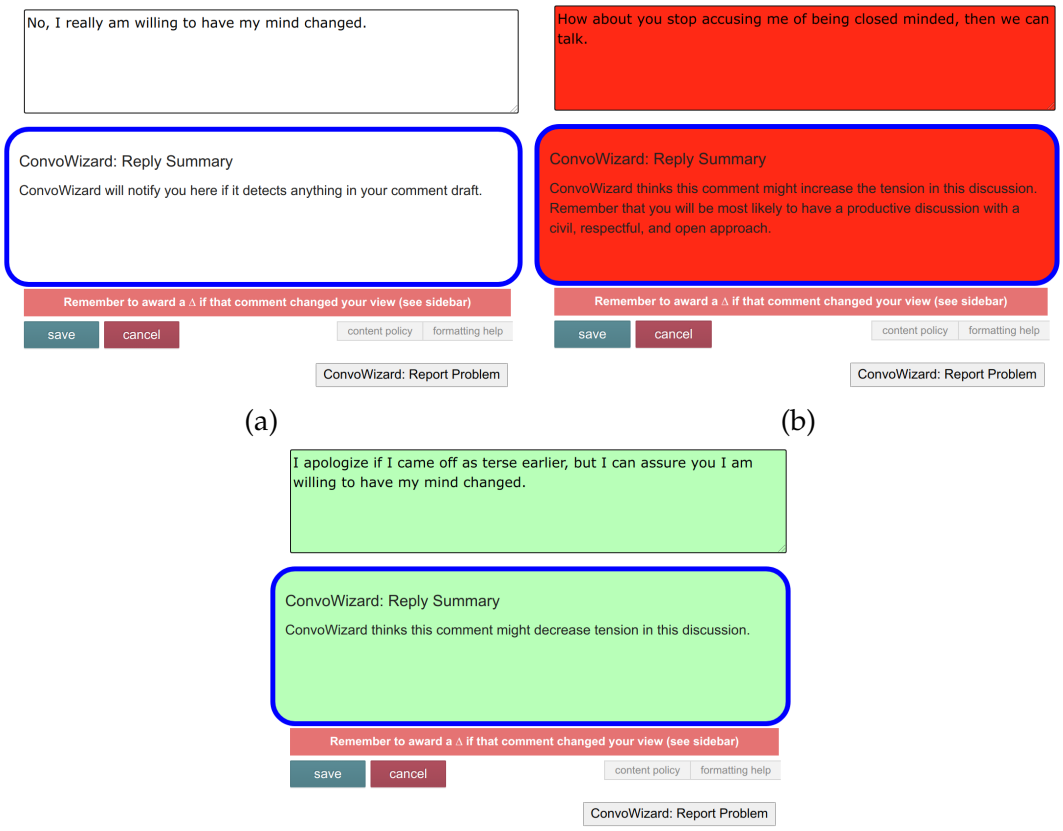


Figure 5.2: The Reply Summary provides information about what impact the user's in-progress draft reply might have on the risk of incivility. (a) If the risk score with the draft reply is the same as the risk score without the draft reply (within a margin of error), the Reply Summary displays a neutral message. (b) If the risk score increases, the Reply Summary displays a warning message with a red background, with deeper shades of red indicating higher resulting risk. (c) If the risk score decreases, the Reply Summary displays a message about decreased tension with a green background, with deeper shades of green indicating larger magnitudes of score decrease. (Note that all three examples shown are replies to the tense context from Figure 5.1c; the preceding context is excluded for readability.)

DOM,[2] and therefore does *not* require any access to the user's Reddit account. The ConvoWizard extension activates whenever a user hits the "reply" button in the Reddit UI, indicating they are considering joining the discussion. As the user drafts their reply, ConvoWizard provides feedback directly inside the Reddit UI via DOM manipulation. It specifically provides two types of feedback, referred to as the Context Summary and the Reply Summary, which are each displayed in separate UI elements (demonstrated in the Video Figure).

The **Context Summary** gives an estimate of how likely the conversation was to turn uncivil *prior* to the user joining in. To produce this estimate, the extension extracts the text of all preexisting comments $\{c_1, \ldots, c_n\}$ in the conversation history from the DOM. Then, it sends this information to the backend server (Section 5.3.1) which returns a CRAFT score $S_{context} = \text{CRAFT}(\{c_1, \ldots, c_n\})$; henceforth we refer to these scores as *risk scores* to emphasize that in the context of ConvoWizard, CRAFT is being used as an estimate of the risk of future incivility. If $S_{context} > 0.55$,[3] the Context Summary displays a warning to the user that the conversation they are about to participate in is tense and might become uncivil in the future. It also visually indicates this risk by changing its background color to a shade of red (scaling by risk score, such that higher scores produce redder colors). This functionality is visualized in Figure 5.1.

Then, as the user drafts their reply, the **Reply Summary** provides real-time estimates of how the in-progress reply, if posted as-is, might impact the risk of the conversation turning uncivil in the future. Every five seconds, the extension sends the current text of the in-progress reply, which we call $r(t)$ (where t represents the current timestamp), to the ConvoWizard backend, which returns a risk

[2]Document Object Model, the browser's internal JavaScript-compatible representation of the web page.

[3]This threshold was determined as part of the original CRAFT experiments (Section 4.6).

score that was computed with this text included: $S_{r(t)} = \text{CRAFT}(\{c_1, \ldots, c_n, r(t)\})$. The Reply Summary then determines what feedback to give by comparing $S_{context}$ and $S_{r(t)}$.[4] If $S_{r(t)} > S_{context}$, the Reply Summary displays a warning that the in-progress reply might increase the tension in the conversation, and visually indicates this with a red background whose shade scales with $S_{r(t)}$. On the other hand, if $S_{r(t)} < S_{context}$ and there was preexisting tension in the conversation (i.e., $S_{context} > 0.55$), the Reply Summary displays a message that the in-progress reply might decrease the tension, and visually indicates this with a green background whose shade scales with $S_{context} - S_{r(t)}$.[5] This functionality is visualized in Figure 5.2.

Backend server

ConvoWizard also consists of a backend server component which is responsible for both running CRAFT to produce risk scores requested by the frontend, and logging the request data to produce a record of how users interacted with ConvoWizard. Every time the backend receives a request from the frontend, it first runs CRAFT on the attached data to produce a risk score to return to the frontend, then it logs the request and response to a database. Each logged request/response object includes the Reddit ID of the comment being replied to, the timestamp t of the request, the generated risk score, and (for Reply Summary requests) the in-progress reply text $r(t)$. Additionally, all requests that happened under the same reply action (i.e., the initial Context Summary request that was sent when the user hit the "reply" button and all subsequent Reply Summary

[4] All comparisons apply a small noise threshold to prevent basing feedback on spurious variance in scores.

[5] The "decreasing tension" intervention is only implemented for conversations with tension in the context because initial testers reported that it was confusing to hear about "decreasing tension" when there was no tension to begin with.

requests until the reply is submitted or cancelled) are grouped together in the database as a single *interaction*. Knowing that a series of requests came from a single interaction allows us to subsequently analyze how users modified their drafts over time, as we will discuss in Section 5.4.

Ethical considerations for technical design

As previously mentioned, the real-world setting of our study raises important ethical challenges. While we primarily respond to these challenges through the design of the study, as we will discuss in Section 5.3.2, there are also ethical implications for the design of the ConvoWizard tool itself.

First, there is the problem of *misuse*: our risk awareness paradigm is designed for well-intentioned users, who do not deliberately desire conflict and are thus more likely to respond appropriately to warnings of potential future incivility. By contrast, bad-faith trolls could respond to such warnings quite differently, for example purposely trying to write a comment that triggers a warning. Thus, it is important to restrict access to ConvoWizard so that bad-faith trolls cannot easily get ahold of it. To this end, ConvoWizard is programmed to be inoperable until it is "activated" using unique credentials that we assign to each participant in our study. To prevent bad-faith trolls from circumventing this restriction by simply signing up for the study, we check the posting history of each user who signs up for our study, and prevent them from joining if they do not have an established history of participation on ChangeMyView (as this might indicate that the account is a purpose-made "sockpuppet" (Kumar et al., 2017) or an outsider seeking to "brigade" the subreddit (Georgakopoulou et al., 2020)).

There is also the problem of *errors* in ConvoWizard's algorithm-driven estimates of risk. Algorithms that operate on human language, and especially on subjective aspects like civility, are far from perfect—they can fail to pick up on nuances of human behavior (Gillespie, 2020; Duarte and Llansó, 2018) and encode biases present in their training data (Davidson et al., 2017). But in the public consciousness, the capabilities and objectivity of algorithms are often overestimated (Bory, 2019; Katzenbach, 2021). To address this we have crafted the messaging in and around the ConvoWizard tool to counteract such possible overestimation by users. Throughout the instructions all study participants must read to set up ConvoWizard, we repeatedly remind them that ConvoWizard is an early prototype and may therefore make mistakes, and we encourage them to report any mistakes they notice.

Furthermore, the warning messages displayed in the ConvoWizard browser extension were specifically crafted to come across as *informational* rather than *prescriptive*—we avoided any wording that might imply the tool is advising users that they should (or should not) post their draft, as well as any language that might be associated with assigning blame. The final wording, which frames ConvoWizard findings as simply the existence or nonexistence of "tension" in the conversation (see Figures 5.1 and 5.2), was decided upon after multiple rounds of internal testing where testers evaluated the messages on whether they contained any of the implications we seek to avoid.

Of course, the steps listed here cannot completely eliminate the possibility of misuse or misinterpretation, and they are not meant as a standalone solution. Rather, these design choices comprise just one step in our broader response to the ethical challenges of this study, which we discuss further in Section 5.3.2.

5.3.2 Study Design

Having developed the ConvoWizard tool as a concrete implementation of the risk awareness paradigm, we now turn to describe the design of our IRB-approved study in which users tested and gave feedback on ConvoWizard in real online discussions.

Community collaboration with ChangeMyView

As previously mentioned, the need to evaluate our proposed paradigm in a real-world setting raises important ethical challenges, due to the danger of harm arising from algorithmic flaws or misuse of the ConvoWizard tool. While we have taken concrete steps to minimize the possibility of harm (Section 5.3.1) such steps can never completely eliminate the possibility.

Any harm that does occur might not just be limited to the users of ConvoWizard—algorithmic flaws or misuse could negatively impact the discussions in which those users partake, and this could have further impacts on the community (subreddit) in which the discussions occur. The resulting ethical implication is clear: the potentially affected party, that is the community itself, must be allowed to play an active role in the setup and execution of the study. This led us to develop our study as a *community collaboration*, actively working together with a specific subreddit and giving its members a chance to weigh in. We specifically chose to collaborate with the subreddit ChangeMyView, a community centered around good-faith debates. We chose this community for two reasons: it has an established history of research collaborations (Jhaver et al.,

2017; Hidey et al., 2017; Wei et al., 2016; Tan et al., 2016),[6] and their overall culture, which prioritizes civility and open-mindedness, is a particularly good fit for our proposed paradigm, which is predicated on the good faith of users.

On Reddit, the term "moderator" can be somewhat misleading—volunteer subreddit moderators are not merely responsible for rule enforcement, but rather play a larger social role as *community leaders*, who engage directly with members of the community both formally and informally to build solidarity and construct shared norms (Dosono and Semaan, 2019; Seering, 2020; Gilbert, 2020) and even serve as their community's representatives to the outside world (Seering et al., 2020).[7] In light of this, our collaboration with ChangeMyView centered around an ongoing dialogue with the ChangeMyView moderators. We first reached out to them to explain our research and propose a collaboration, and after they collectively agreed to the proposal, we worked together to craft a public announcement explaining the study to the broader ChangeMyView community. The moderators subsequently posted the announcement as an official pinned post,[8] which through the course of the study served a dual purpose as both a sign-up hub hosting links to join the study, and as a communications hub where ChangeMyView members (whether participating in the study or not) could ask questions, give feedback, or raise any concerns. As the study proceeded, we maintained our dialogue with the moderators, who acted as intermediaries between us and the ChangeMyView community: they passed along new questions and concerns to us, and we provided them with answers

[6]These collaborations are publicly promoted on the ChangeMyView community wiki: `https://www.reddit.com/r/changemyview/wiki/research`

[7]Concrete examples of this type of work among ChangeMyView moderators include organizing semi-regular town-hall-style feedback threads (`https://www.reddit.com/r/changemyview/wiki/metamondays`) and producing a ChangeMyView podcast (`https://www.reddit.com/r/changemyview/wiki/podcast`).

[8]Official pinned posts always appear at the top of the subreddit page and have special styling to visually distinguish them from regular posts.

and updates which they could add to the pinned post.

Experimental design

In order to determine how users might react to ConvoWizard's interventions, we employ a two-phase user study, consisting of a first phase focused on collecting self reports of how participants use ConvoWizard, followed by a second phase designed to collect more controlled usage data for the sake of quantifying the participant-reported effects. This two-phase design was driven by the aforementioned practical challenge of recruiting regular ChangeMyView users to use ConvoWizard: as we have described, addressing this practical challenge requires that users perceive ConvoWizard as providing real value, and a controlled setup can undermine this since ConvoWizard would not provide any utility to the user within a Control condition. Having two phases offers a workable compromise, as the uncontrolled first phase allows users to experience ConvoWizard in full without having to worry about interference from experimental controls, and serves to ease them in to the more complicated (from the user perspective) controlled second phase.

In Phase 1 of the study, lasting 30 days, participants are asked to install a version of ConvoWizard that does not implement any experimental controls, thus giving all participants an uninterrupted experience of using the tool. The focus of this phase is to gather self reports of how participants interact with this new paradigm, which they provide through an exit survey distributed at the end of the 30-day period (described in more detail later in this section).

Phase 2 of the study, lasting 60 days, is designed quantify the participant-

reported effects from Phase 1 through a controlled analysis of ConvoWizard usage logs. To this end, ConvoWizard in this phase implements a *within-subjects* randomized controlled experiment design in which we assign Treatment and Control conditions at an interaction level: when a participant first hits the "reply" button on a discussion thread within a ChangeMyView post, ConvoWizard randomly decides (with probability 0.5) whether or not to show the interventions for the user's interactions on that post. This way we can compare how each participant behaves in the presence (vs. the absence) of the ConvoWizard intervention. Our choice of a within-subjects design rather than a between-subjects one was again driven by practical considerations: asking users to install and use a tool that does nothing (as would be the case in the Control setting of a between-subjects study) would be infeasible, whereas the within-subjects design allows every participant to experience ConvoWizard's functionality at least some of the time.

Our choice of a within-subjects design rather than a between-subjects one was driven by the aforementioned practical challenge of recruiting regular ChangeMyView users to use ConvoWizard: asking users to install and use a tool that does nothing (as would be the case in the Control setting of a between-subjects study) would be infeasible. By contrast, the within-subjects design allows every participant to experience ConvoWizard's functionality at least some of the time.

For similar practical considerations around recruiting participants, we adopt a two-phase study design. In the first phase, lasting 30 days, there is no Treatment-vs-Control setup, and instead ConvoWizard is always active for all users. While this does not provide controlled data, it does give participants

a chance to experience ConvoWizard uninterrupted, develop an impression of its usefulness, and provide feedback through an exit survey (described below). Participants who indicate in this exit survey that they are interested in a follow-up study are invited to participate in the second phase of the study, which lasts 60 days and implements the randomized controlled experiment design described above.

In total, 47 users finished Phase 1 of the study (including the exit survey) and 14 users finished Phase 2. We acknowledge that this results in a self-selected participant pool that is not necessarily representative of the entire ChangeMyView user population, being more likely to attract users that are interested in the issue of incivility. Despite this limitation, the resulting data can still be useful as a first step towards characterizing the potential of the risk awareness paradigm, as we seek to do in the subsequent analysis (Section 5.4). We return to discuss this limitation—and the steps needed for future work to overcome it—in more detail in Section 5.6.

Exit survey

The exit survey, sent to all Phase 1 participants after the end of the 30-day period, gave participants a chance to report on their experiences with ConvoWizard and provide their overall impressions of the tool, and serves as an instrument for a qualitative evaluation of the risk awareness paradigm. The full text of the survey can be found alongside further details about the execution of the study in Appendix B.

The exit survey contains a mix of multiple-choice questions and open-ended

text responses. It asked specific questions about how participants tended to respond when ConvoWizard warned them about risk of incivility—including whether they tended to agree with ConvoWizard's predictions and whether this subsequently affected their behavior—and also asked more general questions about participants' overall impressions of ConvoWizard and their willingness to use it in their everyday ChangeMyView participation if it were hypothetically available for general use outside the context of the study.

Data collection and processing

As described in Section 5.3.1, ConvoWizard records users' drafting behavior in real time. This data collection takes place for every interaction regardless of whether the tool is in Treatment mode or Control mode, and the result is a rich record of how users draft their comments both "naturally" (in the Control condition) and in the presence of the ConvoWizard intervention (in the Treatment condition). In our subsequent analysis we compare the drafting behavior in these two conditions.

To avoid attributing spurious differences to ConvoWizard, the data must have the following properties:

- Each user should contribute an equal number of Treatment and Control interactions. This prevents our analysis from uncovering spurious differences arising from individual personality traits of the participants.
- The Treatment and Control data should have the same distribution of estimated prior risk. This prevents our analysis from uncovering spurious differences arising from how participants react in discussions with differ-

ent levels of risk.

While with enough participants these properties would follow from the randomization of the experiment design, considering the relatively small number of participants we take an extra step to enforce these properties in our data. For each logged interaction taking place in the Treatment condition (i.e., when ConvoWizard is active), we match it with an interaction from the Control condition that was from the same author and had the same level of estimated prior risk (i.e. context risk score).[9] Any interactions that could not be matched are discarded. This procedure results in a total of 334 pairs (668 total interactions).

5.4 Findings

In order to probe the feasibility of our paradigm, we aim to understand whether informing a user that a discussion they participate in is (algorithmically-inferred to be) at risk of derailment will lead them to attempt to mitigate this risk. Leveraging the mixed methods setup of our study, we address this question by combining qualitative and quantitative insights derived both from exit survey responses and statistical analysis of data collected in the randomized controlled experiment. In survey responses, participants identify key ways in which ConvoWizard's algorithmically-provided risk awareness augments their existing intuitions about risk of derailment. We use these insights to guide an exploratory analysis of how users drafted their comments in the Treatment versus the Control conditions of the randomized controlled experiment. More specifically, the

[9]The matching algorithm prefers Phase 2 data, but is allowed to draw Treatment data from Phase 1 in the rare case where a Control interaction did not have any valid Phase 2 Treatment match meeting both filter criteria.

rest of this section is organized as follows:

1. We investigate whether users judge risk estimates from an (imperfect) algorithm to be a helpful addition to their own intuitions about risk. Survey responses suggest that the answer is yes: users largely find ConvoWizard judgments reasonable, and point out specific situations in which ConvoWizard's warnings helped them identify tension that they might not have picked up on otherwise. As a further promising sign, users express a willingness to use the tool as part of their regular Reddit commenting workflow (Section 5.4.1).
2. Following up on the finding that users judge algorithmic input helpful in deciding when and how to act proactively, we seek to understand in more detail what these algorithmically-guided proactive steps might concretely look like. An initial qualitative picture emerges from freeform survey responses: ConvoWizard's warnings lead users to reflect further on the tension present in the conversation and how their draft reply might affect it, and to revise their draft reply in ways that might mitigate the risk of escalation. Furthermore, a quantitative analysis of participants' comment drafting activity in the randomized controlled experiment reveals effects that, although small, corroborate the aforementioned qualitative findings: compared to the Control condition, during the Treatment condition participants tend to spend more time drafting their comments, make revisions that reduce the algorithmically-estimated risk, and shift their language in ways that roughly correspond to the proactive strategies they reported employing to reduce tension (Section 5.4.2).

5.4.1 Usefulness of Algorithmic Interventions

Our first step in exploring the potential of algorithmic risk awareness interventions is to check whether users actually find such interventions to be helpful additions to their process of reasoning about risk of incivility. To this end, we examine participants' exit survey evaluations of their experience with ConvoWizard, with a particular eye towards how and why they rated its interventions as useful (or not).[10]

We find that participants broadly rated ConvoWizard's interventions as both useful and intuitively correct: 77.1% of participants reported that they found the interventions at least somewhat useful, and 68.1% felt that ConvoWizard's estimates of risk were as good as or better than their own intuition. Furthermore, responses suggest that many participants see acting upon ConvoWizard's warnings as being to their benefit: over half of the participants felt that ConvoWizard's warnings stopped them from engaging in fights with other interlocutors during the experimental period (54.3%), and even prevented them from posting a comment they would have later regretted (54.3%).[11]

To put these numbers in more context, we examine participants' open-ended responses, which shed light on exactly *how* ConvoWizard helped them. In these responses, participants identify a number of ways in which ConvoWizard made

[10] In the survey, mostly-identical versions of the ConvoWizard feedback questions were asked separately for the Context Summary and Reply Summary interventions, to prevent participant confusion. Because the results were broadly similar between the two versions of the questions, to avoid redundancy we will refer in the text to the numbers from the Reply Summary version of the questions (chosen because there are a small handful of questions that were specific to the Reply Summary). Full response numbers for both versions of the questions can be found in Appendix B, Section B.3.

[11] In interpreting these percentages, one should consider that not all participants are expected to be in a situation where they are about to enter a fight or post a regrettable comment during the experimental period.

them more aware of tension in conversations and in their draft replies. Some participants felt that ConvoWizard performed *better* than their own intuition at detecting risk, in that it picked up on cases of tension that they would have missed. **PR18** explains:

> **PR18**: I feel like I don't pay attention to specific triggers programmed into the wizard. Even if my message isn't confrontational the way I say it might have an unintended psychological impact I wouldn't have recognized.

For other participants, even if ConvoWizard was not necessarily better than their own intuition, it served as a second opinion providing clarity in cases where their intuition left them uncertain, as **PR13** found:

> **PR13**: In situations in which I would need more context to see where the discussion is going, ConvoWizard's answer is 'yes' or 'no' while mine is 'I don't know yet', and it's usually right still.

Finally, for some participants ConvoWizard played a somewhat more modest but still impactful role: it served as a prompt to think about tension in cases where they wouldn't have been thinking about it. **PR15** elaborates:

> **PR15**: I don't often care about increasing tension. My objective is generally the discussion, not whether I sound polite or not. ConvoWizard sort of reminds me that I should use maybe different language.

Thus, while individual participants might differ in exactly how they benefited

from ConvoWizard's interventions, on the whole we find that ConvoWizard fills various gaps in their reasoning about tension and thus serves to increase their overall awareness of risk.

Another important factor in judging ConvoWizard's usefulness is participants' willingness to continue using it if it were made widely available. Here, we find that 83.0% of participants expressed at least some interest in adopting ConvoWizard as part of their usual ChangeMyView workflow, if it were publicly deployed. Perhaps more importantly, 63.8% of participants felt that if ConvoWizard were to be broadly adopted by the ChangeMyView community, the net effect would be an *improvement* in discussion quality.

Taken together, these results are a promising initial sign that algorithmic risk awareness interventions can be a valuable tool to help users identify tense conversations. That said, it is just as important to note that as an early prototype, ConvoWizard is far from perfect, and participants also identified specific shortcomings that prevented it from being as useful as it could have been. Most notable among these is the issue of *false positives*: when participants were asked about reasons they might sometimes disagree with ConvoWizard's judgments, false positives were a more commonly cited concern than false negatives, with 61.7% reporting that the former was a common issue they encountered, and only 34.1% reporting the latter. False positives can detract from the overall helpfulness of the tool since too many unwarranted warnings can make the tool seem annoying, as **PR2** explains:

> **PR2**: The "false positive" rate was much higher than the "false negative" rate [...] This was helpful in detecting some things that ought to be rephrased, but slightly annoying at times after several re-edits

of the intended comment.

In the extreme, it could also lead to a boy-who-cried-wolf situation, in which users end up dismissing the tool as just always reporting tension regardless of what is actually happening in the conversation, as **PR37** succinctly puts it:

> **PR37**: It seemed to say everything was in danger of tension

These observations mark an important direction for future work. While ideally tools like ConvoWizard would benefit from improved algorithms that make fewer false positive errors, in light of the fact that the algorithm will never be perfect there is a potential design implication here: future work could look into ways to better trade off precision and recall, or even offer users intuitive ways to adjust this tradeoff to their own preferences.

Another important drawback that participants identified was lack of transparency: 48.9% of participants marked "more transparency" as one of the most important improvements they would want to see in a future iteration of ConvoWizard. The lack of transparency limits ConvoWizard's helpfulness in two key ways. First, as **PR11** explains, it can leave users knowing that a conversation is at risk but not knowing what to do about it:

> **PR11**: I think it needs to get better at walking through why it thinks a thread is hostile and why your reply is. It was often left to me to entirely rethink a statement which seemed to say it was better without explaining why that change helped.

Second, similar to the issue that was raised in the discussion of false positives, seeing the algorithm make apparent mistakes with no explanation as to why

can eventually lead the user to tune out the tool's feedback, a situation that **PR13** identifies:

> **PR13**: Since the reply summary feature flipflopped regularly, I ended up not paying a lot of attention to it. So probably also in cases in which it would have been helpful.

Future implementations should therefore seek to integrate recent developments in explaining algorithmic decisions (as seen with toxicity detection, for instance, in the RECAST system (Wright et al., 2021)) to build algorithmic risk awareness interventions that are more explainable and hence, perhaps, more directly actionable.

Overall, while there is clearly more work needed to help algorithmic risk awareness tools meet their full potential, as a preliminary step the results of our study serve to establish that such tools are at least feasible as a means of increasing users' awareness of risk in conversations. Having established this, we now turn to investigate the implications of this increased awareness; that is, what concrete steps users might take to mitigate risk when it is brought to their attention.

5.4.2 How Users Engage With Algorithmic Interventions

Our exploration of how users engage with the enhanced risk awareness provided by algorithmic interventions is guided by prior work on user-facing interventions aimed at promoting pro-social behavior. Specifically, we focus our attention on two types of concrete proactive steps users might take: spending

extra time to consider and react to ConvoWizard's warnings while writing their comment (Kriplean et al., 2012b), and making (token-level) adjustments to their language use (Seering et al., 2019a). In addition to looking for self-reports of such reactions in the survey responses, we also seek to support any self-reported findings with evidence from the experimental data, by running comparative Control-versus-Treatment analyses at the *interaction* level (that is, on the 668 paired interactions described in Section 5.3.2).

We note, however, that the design of the study imposes several limitations on the comparative analysis: the small sample size restricts us to the use of coarse-grained, simplified metrics and necessarily leads to low-powered results, and the within-subjects setup prevents us from inferring broader behavioral changes beyond how users immediately engage with system interventions. As such, the results should best be understood as highlighting potentially interesting trends in order to guide subsequent work, rather than as being exhaustively conclusive in and of themselves.

Deeper reflection and revision

One basic way users might engage with algorithmic warnings of risk would be to spend more time to consider the tension being pointed out by the algorithm and think about how to reword their comment accordingly. This kind of effect was previously shown in Kriplean et al. (2012b)'s work on the "Reflect" intervention, where users reported taking the time to more deeply consider the comment they were replying to, which the authors speculated would "act to counteract our tendency towards knee-jerk reactions"—precisely the kind of impact we sought to achieve with ConvoWizard.

In open-ended responses, several participants indeed report engaging in such reflection and revision. For instance, **PR26** points out how seeing a warning from ConvoWizard might prompt them to review the conversational context more deeply than they would otherwise:

> **PR26**: I don't always read the entire chain of parent comments so the wizard indicating concern lead me to go back and read the entire chain.

PR9 notes that even though they were aware the algorithm is imperfect, it was good enough to prompt reflection on their own in-progress draft:

> **PR9**: I'm sure its not perfect, but in my case it made me rethink what I type.

Finally, **PR2** explicitly mentions spending extra time rewording their comments:

> **PR2**: I spent a lot of time rephrasing. Often there were phrases that in other contexts could signal increasing tension, but would not in the context I typed.

Since reflection is an inherently subjective process we cannot quantify it directly. We can however check for the existence of the time effect that we would expect to accompany increased reflection (and which **PR2** explicitly calls out). To this end, we compute the mean time spent per logged interaction, stratified both by experimental condition and by whether the interaction was judged to be *at-risk* (i.e., there was enough algorithmically-inferred tension that, for Treatment interactions, a warning was displayed, and for Control interactions, a

		(a) Drafting time (seconds)	(b) Correlation between adjusted timestamp and risk score
At-risk	Control	174.2	0.05^{**}
	Treatment	**189.5**	-0.06^{***}
Not-at-risk	Control	124.3	-0.13^{***}
	Treatment	133.4	-0.06^{**}

Table 5.1: Control-versus-Treatment comparisons of two high-level measures of drafting behavior: (a) Average time spent per interaction, in seconds. **Bolded** Treatment values are significantly ($p < 0.05$, Mann-Whitney test) different from their Control counterparts. (b) Correlations between adjusted timestamp (time in seconds since the start of the interaction) and risk score (as determined by CRAFT). Correlations are measured as Spearman's R and stars indicate significance levels ($^{**}p < 0.01$,$^{***}p < 0.001$).

warning would have been displayed had ConvoWizard been active). If the hypothesized engagement effect exists, we expect that there should be an increase in average time per interaction in the Treatment condition—but importantly, because the hypothesized effect is a response to ConvoWizard's warnings, we expect this difference to exist only in at-risk interactions (since that is the only scenario in which ConvoWizard would display an intervention in the Treatment condition and not display one in the Control condition).

We find this exact effect: in at-risk interactions, there is a significant ($p < 0.05$ via Mann-Whitney test) increase in the mean amount of time spent per interaction in the Treatment condition, and no such change in not-at-risk interactions (Table 5.1a). We further note that this increase cannot simply be explained by differences in the length of the comments; in fact, the average number of words per comment does not differ significantly between the two conditions ($p = 0.21$ via Mann-Whitney test). While we reiterate that this analysis cannot directly measure how much participants actually reflect on their drafts, it at least offers some quantitative corroboration of their self-reports.

As a next step, we want to know how this engagement with the tool might translate into concrete changes to the drafting and revision process. To investigate this, we again start from the open-ended responses: some participants report that they use ConvoWizard's risk intensity feature (i.e., the changing colors indicating levels of estimated risk) as a guide, attempting to revise their comment in a way that produces a less intense color (i.e., lower risk score):

> **PR35**: I kept rewording my reply until it stopped showing up orange.

> **PR9**: Often times if the color changed I would reread what I was saying and see if the response maybe came off the wrong way. Helping me then to reword it.

In the randomized controlled experiment data, if users are actively attempting to reduce the degree of tension displayed by ConvoWizard, as suggested by **PR35** and **PR9**, then in the Treatment condition we would expect to see a gradual decrease in the risk score as a comment gets drafted; that is, we expect an inverse correlation between the risk scores of the intermediate snapshots of a draft and their associated timestamps.

We indeed find (Table 5.1b) that in at-risk interactions in the Treatment condition (i.e., interactions where a warning was displayed), there is a negative correlation between timestamp and risk score.[12] Notably, the corresponding correlation for the *Control* condition is actually *positive*. In other words, in at-risk situations the natural tendency is for risk score to *increase* over time as a

[12]We normalize by the timestamp at which the interaction started (i.e., adjusted timestamp = timestamp − timestamp of first snapshot in this interaction, such that the adjusted timestamp of the first snapshot is always 0).

draft is written, and the introduction of the ConvoWizard intervention actually manages to reverse this natural trend. While the correlations themselves are relatively small in magnitude, it should be noted that a rank-order correlation test is a very coarse metric of the phenomenon being investigated here, since an algorithmic warning and subsequent risk-score-decreasing edit could occur at any point—or even *multiple* points—in the drafting process, so the true relationship may not be monotonic over the entire duration of the interaction. In this sense, it is promising that even such a coarse metric can reveal a significant trend, and this suggests the potential for more sophisticated analyses. For example, a larger study could allow for a more precise analysis considering the exact moment of each warning and the subsequent edits it triggers.

Effects on linguistic strategies

Once a user has reflected on the tension identified by an algorithmic intervention, and revised their comment accordingly, does this end up being echoed in the language of the reply they end up posting? Broadly speaking, participants self-report that this is the case, with 71.4% reporting that ConvoWizard warnings affected the language they used in their replies—but what do these changes specifically consist of?

In exploring this question, we must keep in mind that users are somewhat constrained in the extent to which they can alter their language, since ultimately the goal of the conversation is to have a debate and so users cannot make drastic changes that would alter the semantic content of their comment. As such, to the extent that linguistic change occurs, it is aimed at controlling the *tone* that gets conveyed while preserving semantic meaning, as **PR9** and **PR18** explain:

> **PR9**: I thought of better words I could use maybe words that don't sound like I may be trying to provoke a uncivil response.

> **PR18**: I tended to avoid certain key words that I felt the program picked up on whether or not I was being confrontational. The word "you" or any words with negative connotations could be altered without changing the meat of my messages.

More concretely, participants reported that ConvoWizard warnings led to increases along the same four linguistic strategies for proactively preventing derailment that we previously (Section 2.4) identified: politeness (68.0% of participants who reported any linguistic changes), formality (48.0%), objectivity (44.0%), and question-asking (32.0%).

These results inform our subsequent comparative analysis of linguistic effects in the randomized controlled experiment. As with our earlier analyses, given the limited size of the controlled data, we are necessarily limited in the complexity of the linguistic phenomena we can capture in our analysis. To this end, we adopt a similar strategy to that used by Seering et al. (2019a) in their work on interventions for encouraging prosocial behavior: comparing basic *summary variables* that can be computed as simple functions of tokens and parts-of-speech. Our specific choice of summary variables is inspired by—but not exhaustive of[13]—the strategies that users self-reported employing in order to reduce tension:

[13]Notably, we do not consider politeness, since to the best of our knowledge no trained model exists for ChangeMyView comments and additional labeled data would be needed to train such models (existing politeness models are trained on *requests* extracted from Wikipedia Talk Pages and StackExchange comments (Danescu-Niculescu-Mizil et al., 2013a)).

- **F-factor**: This is a simple measure of *formality* introduced by Heylighen and Dewaele (1999). It is computed as:

$$F = (\text{freq(nouns)} + \text{freq(adjectives)} + \text{freq(prepositions)} + \text{freq(articles)} - \text{freq(pronouns)} - \text{freq(verbs)} - \text{freq(adverbs)} - \text{freq(interjections)} + 1)/2$$

  Where freq() measures the frequency of a given word category in a body of text; that is, a count of words of that type normalized by the total number of words in the text. Because the short length of Reddit comments makes the F-factor somewhat noisy, we use a discretized version of the score, adopting an empirical threshold of 0.44 that was inferred by Heylighen and Dewaele based on an analysis of labeled corpora. F-factor is thus discretized as simply "informal" ($F \leq 0.44$) or "formal" ($F > 0.44$). These discretized scores are compared as "formality rates"; that is, the percentage of all comments that get scored as "formal" within a given set of comments.

- **Categorical-Dynamic Index (CDI)**: This is a score derived from function word counts, with higher values indicating a more analytic and cognitively complex writing style, and lower values indicating more reliance on storytelling and personal narratives (Pennebaker et al., 2014). We use this score as it roughly corresponds to our definition of the *objective-subjective* distinction. It is computed as:

$$\text{CDI} = 0.3 + \text{freq(articles)} + \text{freq(prepositions)} - \text{freq(personalpronouns)} - \text{freq(impersonalpronouns)} - \text{freq(aux.verbs)} - \text{freq(conjunctions)} - \text{freq(adverbs)} - \text{freq(negations)}$$

  We note that the CDI is one of the metrics used for quantifying the effects of prosocial interventions in Seering et al. (2019a).

		Formality rate	CDI (mean)	Question rate
At-risk	Control	81.8%	0.06	15.3%
	Treatment	87.1%	*0.09*	**20.4%**
Not-at-risk	Control	92.8%	0.12	11.6%
	Treatment	92.1%	0.11	14.3%

Table 5.2: Control-versus-Treatment comparisons of three linguistic strategies: formality (measured using the discretized F-factor), the categorical-dynamic index (CDI, used as a rough proxy for objectivity) and the rate of question-asking. **Bolded** Treatment values are significantly ($p < 0.05$) different from their Control counterparts, while *italicized* results indicate an almost-significant trend ($p = 0.07$). Significance is tested using Mann-Whitney for comparison of means, and Fisher's exact test for comparison of rates.

- **Question Rate**: This simply computes what fraction of all sentences within a collection of text are *questions*. While in theory there can be some nuance in what makes a sentence a question, prior computational work on questions found that the simple heuristic of checking for a question mark works remarkably well (Zhang et al., 2017b), and so we adopt this heuristic.

Table 5.2 shows the results of comparing each variable in the Treatment and Control, stratified by whether the interaction was at-risk or not. We find modest but notable differences in the comparisons. Compared to users in Control, users in Treatment ask significantly ($p < 0.05$) more questions. They also appear more likely to write comments that are judged as "formal" (according to the discretized F-factor) and have a higher mean CDI (roughly corresponding to a more analytic, objective writing style); though these latter differences do not reach significance, we still consider them interesting trends worthy of further exploration (with CDI in particular verging on the edge of significance at $p = 0.07$). Like the drafting effects, these effects are only found in at-risk interactions, suggesting that, as expected, they are specific reactions to warnings.

Though these differences are promising signs that the ConvoWizard intervention is having an effect, we must acknowledge that they are relatively small in magnitude. To some degree this is expected, since as explained earlier the goal-oriented nature of ChangeMyView discussions constrains the extent to which users can alter their language. That said, the simplistic nature of the language features being measured here may also play a role in the effect sizes we are observing. In particular, while for the sake of accomodating our limited data we specifically chose lexically-derived features, previously described findings about how users intuitively reason about risk of derailment (Section 2.4) suggest that the most informative linguistic signals of tension and lack thereof, such as tone and making things personal, may not be so easily captured at the lexical level alone. As such, a future larger-scale study could aim to collect enough data to enable analysis using more sophisticated NLP approaches, which could better capture such high-level phenomena—and in the meantime our preliminary results here suggest that linguistic effects are, in fact, a promising target for such continued exploration.

Taken together, these combined qualitative and quantitative findings support a potential mechanism through which the risk awareness paradigm can contribute to more civil online discussions: warnings can lead users to reflect more deeply about the impact their replies have on their conversations and to revise the language of their draft in a way that reduces the risk of derailment. These findings suggest concrete directions for both the design and evaluation of future implementations of the paradigm. From a design perspective, future implementations could explore additional functionality to support the reflection and revision process; for example, using human-readable explanations (as discussed in Section 5.4.1) to guide revisions in a more directly actionable way.

From an evaluation perspective, larger-scale studies are needed to measure the reflection and revision effects in more nuanced and robust ways, including taking a more fine-grained look at the drafting process to capture immediate responses to warnings, using more advanced NLP techniques to capture more abstract changes in language, and running a between-subjects assignment to enable analysis of broader behavioral changes. These future steps would build upon the groundwork established by our current preliminary study, and thereby bring the risk awareness paradigm closer to its full potential.

5.5 Risk Awareness Paradigm for Moderators

Our ConvoWizard user study has focused on how forecasting algorithms can help *ordinary users* of online platforms with the challenging task of identifying conversations that are at risk of derailing. However, as we have established in Chapter 2, ordinary users are not the only people who encounter this challenge: *moderators* report similar difficulties in finding, and keeping track of, at-risk conversations. Given that the results of the user study show promising indicators that ConvoWizard and its risk awareness paradigm might concretely benefit users, a natural follow-up question is whether a similar approach could be applied to help moderators.

Fully evaluating this question would require developing a moderator-facing tool and testing it with a user study, like what was done for users with ConvoWizard. While such a user study does not yet exist, we wish to lay the groundwork for such a study by exploring what moderators might think of algorithmically-assisted proactive moderation at a conceptual level. To this end,

we now describe the design of a prototype moderator tool based on the same technology that powers ConvoWizard, as well as moderators' initial reaction to this proof of concept.

5.5.1 Prototype Tool for Assisting Proactive Moderation

Our prototype tool is implemented as a password protected website that includes two main features: a ranked view of ongoing conversations ordered according to their likelihood of derailing into future antisocial behavior (Figure 5.3), and a conversation view giving a comment-by-comment breakdown of risk levels within the discussion (Figure 5.4). The tool currently works on both ChangeMyView and Wikipedia Talk Pages (discussed in Section 2.3.1), though for the sake of this discussion we focus on the latter setting, as this enables us to gather initial reactions as part of our interviews with Wikipedia discussion moderators (Section 2.3.2).[14]

Frontend: The Moderator's Display

Our prototype frontend consists of two sections: a Ranking View and a Conversation View. The frontend adopts design metaphors used in existing Wikipedia moderation tools, and comes with a broad range of features and parameters in order to engage interview participants in a discussion that can inform future design.

The "main page" of our prototype tool's frontend interface is the **Ranking**

[14] A video demonstration can be found at `https://www.cs.cornell.edu/~cristian/Proactive_Moderation.html`.

Top 78 Conversations Sorted by Score, High to Low:

Rank	CRAFT Score	Score Change	Conversation Name
1	0.776	↘	Kim_Jong-un: fix the romanization of Kim Jong-Un and other Korean names.
2	0.619	↗	Donald_Trump: donald trump is big gay
3	0.54	↗	Global_warming: Whitewashing alert
4	0.534	→	Donald_Trump: Piss blackmail allegations
5	0.512	↗	Kim_Jong-un: Infobox
6	0.507	↘	Bernie_Sanders: "Honeymoon" trip to Russia
7	0.472	↗	Donald_Trump: Highlighted open discussions
8	0.439	→	Donald_Trump: The Current Consensuses
9	0.438	→	Barack_Obama: Doublespeak of equating bombing with activism offensive
10	0.425	→	Global_warming: 2020 research - cognitive difference due to terminology has vanished
11	0.416	↘	Global_warming: HURR
12	0.415	→	Donald_Trump: New NEWS today, for future editing
13	0.398	↗	Bernie_Sanders: how come his praise of socialist regimes are missing from this detailed article?
14	0.386	↘	Donald_Trump: headline
15	0.386	↘	Global_warming: Survey

Figure 5.3: The *Ranking View* of our prototype tool, showing a list of live conversations on Talk Pages, sorted by their predicted risk of derailing into antisocial behavior.

View (Figure 5.3), which is inspired by the organizational concept of a *work queue*—a common interface among existing Wikipedia tools (Halfaker et al., 2013; Geiger and Ribes, 2010). Based on a list of Talk Pages to include, our prototype tool provides ranked list of all ongoing conversations on those pages, sorted in the order of their CRAFT-estimated risk of derailment. CRAFT scores are computed based on all the comments posted so far in the conversation, i.e., CRAFT($\{c_1, c_2, \ldots, c_n\}$). Additional visual cues are implemented to help moderators quickly identify the riskiest situations at a glance: conversations in the ranking are color coded according to their CRAFT score (higher score being deeper red), and each conversation in the ranking is additionally decorated with an arrow whose direction and size reflect the gradient and size of the most recent

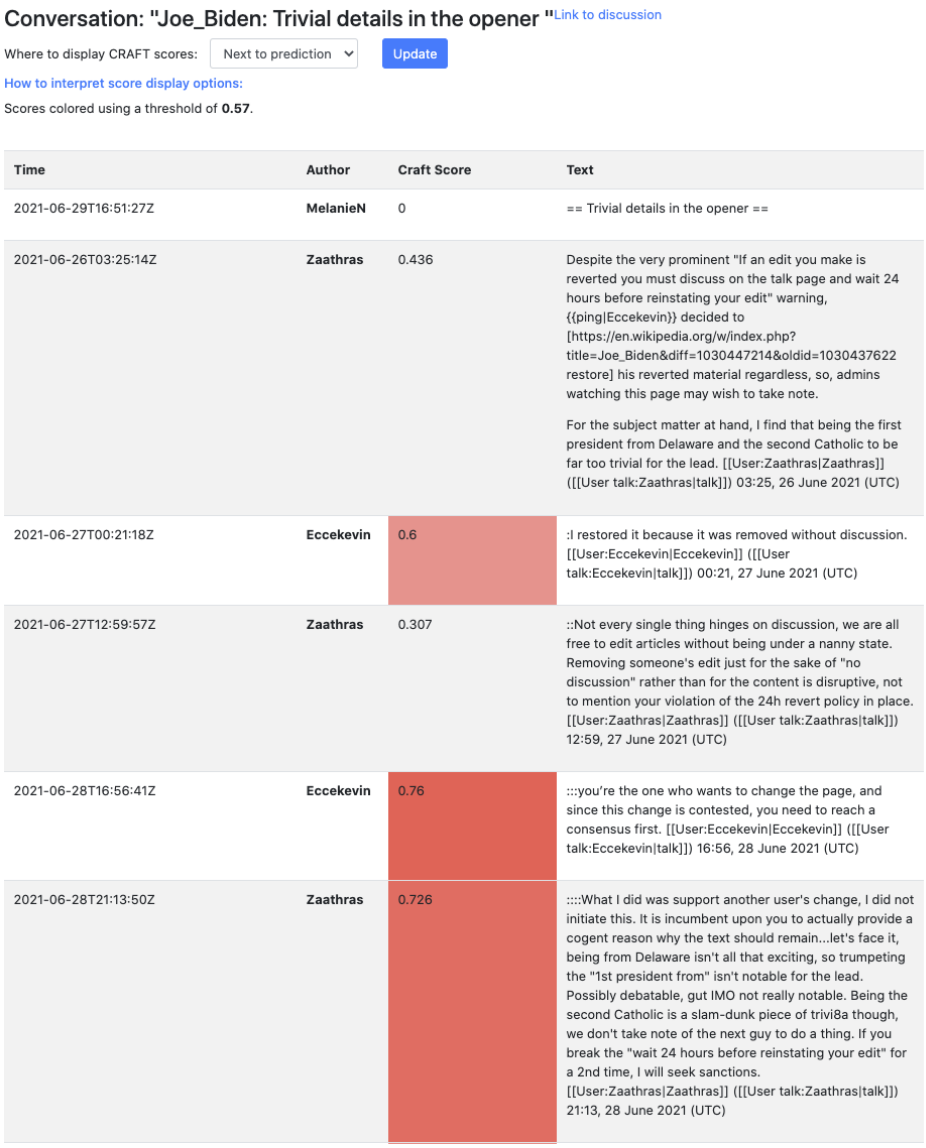


Figure 5.4: The *Conversation View* of our prototype tool, showing a conversation with CRAFT scores alongside each comment. Each score represents the predicted risk of derailment at the time the corresponding comment was posted (taking into account the entire preceding context).

change in CRAFT forecast, i.e., $\text{CRAFT}(\{c_1, c_2, \ldots, c_n\}) - \text{CRAFT}(\{c_1, c_2, \ldots, c_{n-1}\})$. The latter visual cue is meant to help identify rapidly escalating situations; for example a large red up-facing arrow would signal that tension is rapidly rising.

In addition to displaying summary level information about a conversation,

each row of the Ranking view is a clickable link that leads to the **Conversation View** (Figure 5.4), which displays the entire history of that conversation. The Conversation View presents the text of each comment in the conversation along with the time it was posted, its author, and the CRAFT score (color coded as before) at the time that comment was posted, i.e., taking into account the conversation up to and including that comment. This provides some level of transparency as to why the algorithm placed the conversation at a certain position in the Ranking View, allowing the moderator to observe how the predicted risk evolves as a conversation progresses. This design bears similarities to how algorithm decisions are presented to moderators in Crossmod (Chandrasekharan et al., 2019), an experimental tool for assisting (reactive) moderation on Reddit.

Backend server

Our prototype tool's backend server is similar to the one used by ConvoWizard (Section 5.3.1). Like the ConvoWizard backend, it is responsible for running CRAFT to produce risk scores to return to the frontend. The main difference is that since moderators may want a broader view of discussions throughout Wikipedia, not just ones they have participated in, this version of the backend server must automatically keep track of all live conversations on a selected set of Talk Pages.[15] At regular intervals, the backend pulls the latest updates to every Talk Page being tracked, parses the updates to extract the conversations happening on the page, and runs CRAFT on those conversations to get an updated forecast of the risk of future incivility for each con-

[15]In practice we expect that a production-ready tool would need to let moderators choose which pages they want to monitor, but for the sake of this proof of concept we hard-coded a set of pages that we reasoned are likely to have conflict and need moderation: Barack_Obama, Bernie_Sanders, Coronavirus_disease_2019, COVID-19_pandemic, Donald_Trump, Joe_Biden, Kim_Jong-un, and Global_warming.

versation. The tool also keeps track of how the CRAFT forecast for a discussion has changed over time. That is, for a (possibly ongoing) discussion $D = \{c_1, c_2, c_3, \dots\}$, the tool creates and maintains a history of CRAFT forecasts $\{\text{CRAFT}(\{c_1\}), \text{CRAFT}(\{c_1, c_2\}), \text{CRAFT}(\{c_1, c_2, c_3\}), \dots\}$.

5.5.2 Moderator Reactions to the Prototype Tool

To gather some initial insights into the usefulness of algorithmically-assisted proactive moderation, we showed our prototype tool to Wikipedia discussion moderators and asked them for open-ended feedback. This took place as part of the broader process of moderator interviews discussed in Section 2.3.2. We note that as this process can only surface high-level qualitative feedback, these results are not meant as a substitute for a user study, but instead meant as preliminary insights that can guide the design of such a study.

Moderators' feedback on the prototype tool suggests that information presented in the tool's *Ranking View* is helpful in discovering at-risk conversations, although individual moderators differed in their evaluation of exactly *which* pieces of information were most useful. For example, **PW4** reported that they would mainly use the CRAFT score to decide which conversations were worth monitoring:

> **PW4**: [For monitoring] I would just pick the ones with the highest score 'cause it seems to be somewhat accurate.

Meanwhile, other participants highlighted the score change representation (i.e., the colored arrows) as providing an easy way to get a sense of when a monitored

conversation needs to be further inspected. **PW7** reports:

> **PW7**: I like the score change indicator. That is useful. From a cursory glance, it looks like if the score is going up, I would inspect it, if the score was going down, maybe it is not worth inspecting.

All together, five participants described how both the score and score change representation would be useful towards discovering these at-risk conversations.

However, moderators also identified several aspects of conversations that play into their existing intuitions about whether to monitor a conversation, but are not captured by the prototype tool. Some suggestions that were brought up included showing how long a conversation has been active and providing a summary of recent comments in each conversation—suggestions that are worth exploring in a future implementation of this tool for a user study. On the other hand, five participants reported wanting to see data about discussion participants such as their usernames or age on the platform or their prior activity—features that could raise practical and moral concerns and whose inclusion should thus be carefully weighed.

The feedback discussed thus far suggests that moderators would find the Ranking View useful in *identifying* conversations that might be at risk. However, as discussed in Section 2.4.3, an important additional part of the proactive moderation workflow is continuing to *monitor* such conversations. While we believe the comment-by-comment information given by the Conversation View could be helpful for this,[16] that would only be the case if this information aligns

[16]In addition to just providing an augmented interface to follow the unfolding conversation, in a future iteration of the tool we can envision additional affordances, such as allowing the moderator to request notifications based on specific CRAFT thresholds.

with how the moderator would intuitively judge the conversation.

To assess this, we selected several conversations from different positions in the ranking and invited the moderators to first examine them raw (i.e., without added information), allowing them to make intuitive judgments, and then to re-examine them in the Conversation View. Overall, moderators reported that the displayed per-comment CRAFT scores matched their own initial intuitive judgments. For instance, while looking at an example conversation predicted to be heading for future toxicity, **PW2** describes:

> **PW2**: [The escalating comment] definitely took it to a whole new level—and then having the third person come in, right? So, I feel like [the conversation view] is backing up what intuitively I had said. [...] I feel like that's very much in line with my experience and makes a lot of sense.

The most notable exception is that some participants disagreed with the final CRAFT score of a conversation because they thought the conversation was unlikely to continue, and thus in a trivial sense unlikely to see any future toxicity. **PW8** explains:

> **PW8**: I didn't think [the last comment has] that high [chance of seeing a toxic reply]. I mean, in most cases this person [...] will rage quit. That's typically in my experience what happened. That's interesting. I didn't think it was going to be that high [of a score].

This suggestion points to the importance of considering outcomes beyond just future toxicity—something which could be explored in future work not just in

moderator-facing tools, but also user-facing tools like ConvoWizard.

5.5.3 Implications for Future User Studies

Taking the design implications of a proactive moderation tool gleaned from moderators' feedback, together with the observation that CRAFT's forecasts generally agree with moderators' intuitions, we conclude that it is at least feasible to support moderators in identifying and monitoring at-risk conversations. However, this conclusion does not necessarily imply that moderators would accept and use such a tool. **PW3** explains some hesitations:

> **PW3**: I think an algorithm could be a useful indicator for flagging, 'Hey, this seems like a topic or a conversation that might be a problem down the line.' But on its own I don't think an algorithm could actually be trusted to make the decision. A nice little browser plugin that highlights a section for me that says, 'This discussion looks like it's getting heated, you might want to take a look at it,' that's something I would trust. A browser plug in telling me or a pile of machine learning telling me, 'Block this person, they're making everything uncivil wherever they go,' not as inclined to trust it.

As **PW3** exemplifies, moderators are rightfully hesitant to put their full faith in an algorithmic tool, preferring to only use such a tool under their watch. Therefore, despite the agreement between state of the art forecasting methods and moderators' intuitions, these considerations motivate the need for follow-up work to conduct a large scale user study to more systematically analyze how moderators would use such a tool.

In addition to exploring the technical design choices of what parts of the proactive moderation workflow an algorithmic tool would handle on its own versus what parts it would defer to human moderators, and quantitatively measuring how such a tool might impact human moderators' workflow, a hypothetical follow-up user study would also need to be conscious of several ethical concerns that are specific to the moderator-facing setting. One such consideration is being careful about how existing proactive moderation practices, while being generally accepted when done by human moderators, might veer into questionable territory when scaled up by algorithmic tools. In particular, while some moderators reported that knowing *who* is in a discussion plays a role in their intuition about whether the discussion is at risk (i.e., because two users in the discussion are known to get into fights often), incorporating such information into an algorithmic tool might be considered as algorithmic profiling, which can be problematic. **PW3** succinctly expresses this dichotomy:

> **PW3**: I think one thing that actually could be potentially useful for this is, though it also gets into some questionable territory is: who is in the discussion. Either just a breakdown of the top five contributors to the discussion. Or even, if we want to go into more Big Brother territory, [a summary of] how this person's comments usually score.

More broadly, the fact that moderators are inherently in a position of authority, with the ability to broadly influence outcomes in their communities, amplifies the risk of harm from algorithmic errors; for instance, a bad moderation decision that was influenced by flawed algorithmic feedback could negatively impact trust in moderators and thereby have negative repercussions beyond a single user or thread. This implies the need for additional safety guardrails,

on top of the ones already implemented in the ConvoWizard user study. One promising approach to safely conducting user studies with moderators, as seen in prior work (Chandrasekharan et al., 2019), is to run the studies in sandboxed environments that draw their data from real discussions but are kept separate from the actual public-facing community.

5.6 Discussion

Throughout this dissertation, we have operated from the viewpoint that the solution to toxicity in online discussions should come, in part, from the participants in these discussions. They can—and, as they indicate in our surveys, do—use their conversational skills to proactively reduce tension when they are aware that the discussions they engage in may be at risk of derailing into uncivil behavior. However, they sometimes also miss the opportunity to react and use these prosocial skills, in which case they may end up escalating the tension or even reply with an uncivil comment they later regret posting.

Starting from this premise, this chapter has proposed a new proactive paradigm which seeks to prompt participants to employ their prosocial conversational skills by enhancing their awareness about the risks of the discussions they engage in. To demonstrate the potential this paradigm has in a real world setting, we developed an algorithmic tool that can inform a user about existing tension in their conversation and in their reply draft in real time, and conducted a user study in a popular debate community. The results show that users are indeed responsive to the additional risk awareness provided by our tool: the tool's warnings prompt participants to spend more time (re)considering their

language, and activate conversational skills that they normally employ to reduce tension in conversations.

Unlike solutions that rely solely on moderators, the risk awareness paradigm is decentralized and thus can more easily scale with the number of users on the platform. As such, tools based on this paradigm could be a valuable addition to the broader arsenal of moderation strategies employed by online communities. However, fully deploying such tools at scale requires first carefully understanding the impacts they might have on users and the community as a whole. Our present work takes an important first step towards this understanding, using a small-scale study to establish the necessary groundwork for subsequent larger scale follow-ups and identify specific directions that such future work should pursue more deeply, as we discuss below.

Model error and ethical considerations. Any tools interfering in online discourse through algorithmic means should be subject to ethical scrutiny. Unlike paradigms that seek to outright automate the moderation process, our approach aims to merely provide information to the users, and does not trigger harsh actions such as content removal or user banning. Nevertheless, tools like ConvoWizard still have an inherent potential for negative consequences due to their reliance on imperfect algorithms—giving users erroneous information about the risk level in their conversations could cause harm, especially if these errors arise from model bias against marginalized groups.

It must further be noted that even in the absence of model error, there are still ethical concerns at a more conceptual level. While the risk awareness paradigm aims to improve the civility of online discourse, "civility" is ill-defined and often varies by community (Chandrasekharan et al., 2018), and there can be a fine

line between incivility and mere disagreement (Arazy et al., 2013). As such, the risk awareness paradigm—like other moderation strategies—may risk creating a chilling effect on speech that disincentivizes users from expressing disagreement at all (Gillespie et al., 2020) or "tone policing" the type of disagreement that does end up happening, restricting free expression in a way that might systematically silence certain social groups (Gorwa et al., 2020). These concerns are exacerbated by the observation that the lines between incivility and disagreement are especially likely to get blurred in debates over contentious or controversial topics (Crawford and Gillespie, 2016), which are exactly the cases where it is particularly important to make sure that already-marginalized voices are not further silenced.

We have been cognizant of these potential harms in designing our study, and the need to account for them ended up shaping key parts of our study design, such as purposely avoiding prescriptive and blame-assigning language (Section 5.3.1) and running our study as a collaborative effort with community input (Section 5.3.2). However, further work is needed both to more rigorously characterize the potential harms that can arise from erroneous risk level estimates, and to explore further ways of mitigating these harms. In particular, future work should look into ways to make algorithmic risk awareness interventions more *transparent* and *explainable*, which could shed light on algorithmic biases and help users make more informed decisions about each individual intervention (Wright et al., 2021).

Well-intentioned users. As we have previously described, our proposed risk awareness paradigm is designed to be used by well-intentioned users—that is, those "ordinary" users who seek to engage with and contribute to their com-

munity in good faith, as opposed to deliberately seeking conflict, and who comprise the majority of users within many communities including ChangeMyView. While our exit survey results suggest that participants in our study meet this description, we must acknowledge that self-selection effects likely resulted in a participant pool that is not necessarily representative of well-intentioned users in general; specifically, users who are willing to volunteer for a study on civility may do so because they are unusually thoughtful about civility compared to the average well-intentioned user. In order to move beyond the proof-of-concept stage, future work would need to look into ethically viable ways to scale up testing and evaluate the effectiveness of tools like ConvoWizard in the hands of a more general pool of users who, while still well-intentioned in the sense of not being bad actors, may be less deliberately reflective of tension compared to the participants in our small, self-selecting pool.

Beyond study limitations, a separate concern regarding our risk awareness paradigm's reliance on well-intentioned users might arise when thinking about possible future real-world deployment. While we have argued that most users are well-intentioned, bad actors exist in any community and can misuse publicly available moderation tools towards malicious ends (Jhaver et al., 2019a). A public deployment of a tool like ConvoWizard would likewise be vulnerable to misuse; for example, as described in Section 5.3.1, a bad faith troll could deliberately attempt to craft a message that triggers a warning.

One initial response to this concern is to point out that a similar premise of good faith underlies a number of user-facing moderation tools that already see widespread, large scale use—for example, both community voting (Lampe and Resnick, 2004; Mamykina et al., 2011) and flagging/reporting systems (Craw-

ford and Gillespie, 2016) only work to counteract incivility if they are used by users who actually desire civility, and are theoretically vulnerable to abuse by bad-faith users (Richterich, 2014). This has not stopped such systems from becoming a common part of platforms' moderation toolboxes—they are simply not the *only* tools in those toolboxes (Seering, 2020). We similarly envision tools like ConvoWizard being integrated into a broader moderation ecosystem, which could provide ways of establishing checks and balances against misuse. In particular we expect that moderators—who are best positioned to determine what "well-intentioned" means in the context of their community—could retain a degree of control over the deployment of these tools, in a similar way to how we controlled access to the ConvoWizard prototype to minimize the potential of misuse within the context of the study. For instance, moderators may decide whether risk awareness tools are a good fit for their community at all (as we will further discuss below), or even take a finer-grained approach and set limits on who can access the tool, perhaps using hand-written rules and heuristics (e.g., a minimum activity filter similar to the one we implemented in our study recruitment) in a system like Reddit AutoModerator (Jhaver et al., 2019b). In light of this, a natural next step for future work might be to conduct a study with moderators to get insights on how they might manage the deployment of tools like ConvoWizard, and what concrete features would need to be implemented to meet their use case.

Downstream effects. This work has characterized the effect of ConvoWizard's warnings on how its users draft their replies. But a reply does not exist in a vacuum—it is part of a larger discussion, and so a change in the language of one reply might have further downstream effects on subsequent replies and on the outcome of the discussion. Future work should investigate such down-

stream effects, with a particular eye on whether the pro-social changes triggered by a ConvoWizard warning (Section 5.4.2) might further translate to more civil behavior of other interlocutors (Bao et al., 2021), or whether they strengthen or weaken the persuasive effectiveness of the argument (Tan et al., 2016). An even larger scale study could additionally examine community-level effects, looking for empirical support of participants' self-reported belief that wide adoption of a tool like ConvoWizard would improve the quality of discourse in the community (Section 5.4.1).

Further domains and use cases. Our study has focused on one community, ChangeMyView, which was specifically selected because it aims to host good faith debates (Tan et al., 2016). This naturally leads to questions about how well a tool like ConvoWizard would generalize to other communities. Given the aforementioned targeting of well-intentioned users, it is fair to acknowledge that our paradigm has little value in communities where such users are sparse. Nevertheless, we believe that there are other communities with similar values to ChangeMyView where the risk awareness paradigm could be very impactful. In particular, *goal-oriented* communities, including Q&A communities like StackOverflow and Quora (Mamykina et al., 2011) as well as work-coordination settings like Wikipedia Talk Pages (Wulczyn et al., 2017; Kittur and Kraut, 2008), have an added incentive to keep discussions civil since incivility can distract from their broader non-conversational goals (Arazy et al., 2013). Future work could conduct follow-up studies on such platforms to better understand how the specific needs of these communities might differ from those of debate-centric communities like ChangeMyView, and what implications these community-specific needs might have on the implementation and effectiveness of the risk awareness paradigm.

In addition to exploring other communities, another natural follow-up question is to explore other use cases of the underlying conversational forecasting technology beyond just giving feedback to individual users during the conversation. One possible modification of the ConvoWizard concept might be a tool that gives *public* feedback rather than providing warnings to individual users. From a technical perspective, such a tool could be implemented by connecting the ConvoWizard backend to an automated bot (like those commonly found on both Wikipedia and Reddit) that posts a public reply to at-risk discussions. A follow-up user study could compare and contrast the benefits of this bot-based approach versus ConvoWizard's browser extension approach, especially with regards to downstream effects (as defined above), which might be more prominent in a situation where *everyone* knows about the rising tension. Yet another use case, as discussed in Section 5.5, is to apply forecasting algorithms to help moderators. Future work should expand on the prototype moderator tool introduced in Section 5.5, incorporating the preliminary feedback from moderator interviews, and conduct a user study to quantitatively examine the effects of such a tool much as we have done with ConvoWizard.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

This dissertation started from the premise that online community governance is more than just the stereotypical notion of moderators removing offensive comments: it is a process that involves a large amount of less visible but equally important proactive work, both on the part of moderators and ordinary users alike (Chapter 1). To gain more insight into this proactive work, we interviewed moderators and users, and found that proactively preventing toxicity involves intuitively identifying conversations that are at risk of derailing into toxicity—a task which moderators and users alike find possible but challenging, which we argued represents a potential opening for algorithmic assistance (Chapter 2). As a first step towards making such algorithmic assistance a reality, we established the need for computational models that can perform the novel task of *forecasting conversational derailment*, showed the feasibility of approaching this task computationally (Chapter 3), and introduced CRAFT, a first-of-its-kind model for doing this task practically, in real time (Chapter 4). Finally, we used CRAFT as the engine for a prototype user-facing tool, providing the first-ever concrete evidence of the potential for algorithmically-generated proactive interventions to help users avoid derailment (Chapter 5).

6.1 Our Vision: Forecasting, Computational Tools, and Society

It is also useful, at this point, to step back and revisit the broader motivations for pursuing this work. Online toxicity—and antisocial behavior more broadly construed, including trolling, cyberbullying, and hate speech—is widely recog-

nized as one of the biggest problems facing the social Web, not only within the professional circles of academia (Jurgens et al., 2019; Gillespie, 2018) and law and policy (Gorwa, 2019; U.S. House of Representatives, 2019), but also within the popular media (Marantz, 2018; Newitz, 2020; Brooks, 2022; Cross, 2023). Yet widespread recognition of the problem has not necessarily translated into broad consensus on how to handle it. The reason for such lack of consensus is that online community governance and content moderation aim to balance multiple, sometimes conflicting goals of public interest. While many have legitimate concerns about the harms of online toxicity, so too do many have equally legitimate concerns about the excess power wielded by technology companies (Gillespie et al., 2020), the psychological harms associated with moderation work (Roberts, 2014; Dosono and Semaan, 2019), and the continued ability of the internet to support free expression (Carmi, 2019).

It is evident, then, that there cannot be a one-size-fits-all approach to online community governance, and this explains the vast diversity of approaches to the problem—varying along dimensions of **who** is involved, **what** actions they take, and **when** they take them—that we covered in Chapter 1. Accordingly, we believe that the full potential of algorithmic assistance for community governance (Seering et al., 2019b; Wright, 2022) can only be met if communities' algorithmic "toolboxes" are as diverse as their needs, with different tools geared towards different modes of governance with different aims. In this vision of the future, forecasting-based tools like ConvoWizard will exist alongside other computational tools like toxicity detection algorithms, content filters, and user blocklists, together constituting a thriving ecosystem of methods that work together to improve the experiences of users in online communities. And our results thus far suggest that forecasting-based tools do have a promising place in

this ecosystem, having the potential to help users avoid making comments they might have regretted, and to do so at scale. Yet we are also left with a number of open questions, which suggest several paths forward for future work to more fully characterize the role of forecasting-based tools in online communities and gain an expanded understanding of their overall impacts.

6.2 Future Directions

6.2.1 Improving Transparency and Explainability

As noted in Section 5.4.1, greater transparency was one of the most commonly requested improvements among participants in the ConvoWizard user study. Their freeform responses highlighted how lack of transparency poses a barrier to achieving the full benefits of ConvoWizard: while seeing a warning from ConvoWizard might indeed help users become more aware of the risk of derailment, it is not always easy to *act* upon this awareness if one does not know *why* the risk is high and what can concretely be done to change this. In other words, knowing about tension in the conversation but being unable to act on this knowledge is perhaps not much better than not knowing about the tension in the first place.

Yet as desirable as transparency may be in the abstract, actually achieving it is no trivial matter. Algorithmic explainability is famously an open problem in artificial intelligence and machine learning, birthing an entire subfield known as *explainable artificial intelligence*, or *XAI* (Doshi-Velez and Kim, 2017; Gilpin et al., 2018; Barredo Arrieta et al., 2020). While recent years have seen a number of

advances in XAI (see Adadi and Berrada (2018) for a survey), existing methods remain imperfect and do not provide a full view into the inner workings of so-called "black box" models, which include neural network models such as CRAFT. These drawbacks have led to questions about whether the degree of transparency offered by XAI methods is sufficient for real-world applications, especially ones involving high-stakes decisions (Ghassemi et al., 2021).

To further complicate the matter, achieving transparency in a practical setting involves not only the aforementioned *technical* challenge of making black-box algorithms explainable, but also a *design* challenge of how to present these explanations to users in an actionable way. User studies have suggested that presenting algorithmic explanations to users to aid in a downstream task may not always result in meaningful improvements of those users' performance in the task—and more importantly, that different *kinds* of explanations may be more or less helpful to users (Joshi et al., 2023).

Bearing all these nuances in mind, we now turn to discuss some specific recent advances in XAI—both on the technical side and system design side—and how they might be adapted to the forecasting setting.

The Technical Question: How to Explain Forecasts?

Broadly speaking, technical approaches to explain algorithmic decisions fall into two categories: ones that change the model itself to include human-interpretable components, and ones that operate on a post-hoc basis by analyzing the model's output without directly modifying it. We believe that recent advances in generative AI and large language models (LLMs) can enable both

approaches to be applied to the conversational forecasting setting.

To create forecasting algorithms with built-in explainability, we may consider adapting the recent breakthrough in *Text Bottleneck Models* (TBMs) (Ludan et al., 2023). Inspired by Concept Bottleneck Models (CBMs) from the field of computer vision (Yang et al., 2023), TBMs operate by using off-the-shelf LLM summarization models to identify human-interpretable concepts related to the final classification target, then adding an extra layer—a *bottleneck*—prior to the classification output, in which the model must first predict these bottleneck concepts; those predictions are then fed as input to the final classification layer. A key missing ingredient in making TBMs applicable to conversational forecasting is a method for summarizing conversations in a way that captures the relevant conversational dynamics (see Section 4.1). But recent breakthroughs in LLM-based summarization of conversational dynamics (Hua et al., 2024) might provide this missing ingredient, and we believe it is worth exploring whether this new summarization method can be leveraged to generate useful and interpretable bottleneck concepts for a TBM-based forecasting algorithm to use.

An alternative avenue of exploration, which does not require designing new forecasting algorithms, is post-hoc analysis of an existing forecasting algorithm's predictions. One approach which has shown promise in other domains is *counterfactual explanations*: by systematically making perturbations to the input data (representing counterfactuals about how the data "could have been"), it is possible to gain insight into the model's decision-making process by analyzing what kinds of perturbations made the biggest difference (Wachter et al., 2018). While this approach has shown success in settings like financial applications where inputs often consist of discrete features (e.g., a person's age, gender,

and nationality) that can be independently perturbed, applying it to conversational data poses a bigger challenge, as peturbations will need to be carefully designed to ensure the modified text is still fluent and naturalistic (Fu et al., 2020). Here, again, we believe that LLMs and generative AI offer one possible path forward, as LLMs have shown promising performance in both conversational turn generation (Wu et al., 2023) and paraphrasing of existing text (Witteveen and Andrews, 2019). Applying LLMs to generate counterfactual continuations of a conversation, for the sake of post-hoc analysis as counterfactual explanations have been applied to other fields, is therefore another promising path forward for achieving explainable forecasting.

The Design Question: How to Present Explanations to Users?

Besides the aforementioned technical questions about how to make forecasting algorithms explainable, there is an accompanying design/HCI question about how and when such explanations should be shown to users. The explanations provided by many cutting-edge techniques tend to be highly technical, and simply showing statistics like a feature importance score or a change in probability may be inaccessible to general audiences (Abdul et al., 2018; Zhu et al., 2018). Once again, conversational settings add an extra layer of complexity: because conversations play out over extended periods of time and involve constantly-evolving dynamics, different points in a conversation may be more or less appropriate moments for intervention, and paying attention to the "rhythm" of a conversation might allow for interventions to be reserved only for the most critical moments (Janiszewski et al., 2021; Sicilia et al., 2024). Future user studies will need to more thoroughly investigate how users interact with interventions

in order to identify what aspects of an explanation users find most actionable, similar to work that has been done with traditional toxicity detection algorithms (Wright et al., 2021).

6.2.2 Beyond Language: Incorporating Social Knowledge

While we have thus far focused on *linguistic* factors underlying conversational derailment, participants in our interviews (Chapter 2) also identified several non-linguistic aspects of their intuitions about derailment. This was most common among moderators, who pointed to the fact that over time they get to know users in their community and get a feeling for who does not get along with who, but ordinary users also sometimes pointed to similar insights as well as intuitions about a commenter's underlying intentions and whether they are acting in good faith. These findings point to an important property of derailment: it is not solely a function of the language being used in the conversation, and at the level of human intuition, knowledge of *social factors* can both modify the meaning of the language being used and add additional information beyond the language itself.

These qualitative observations about the importance of social factors in toxicity and derailment are backed up by prior quantitative findings as well. For instance, Danescu-Niculescu-Mizil et al. (2013b) showed that a user's willingness to adhere to community norms is influenced by how long they have been in the community, while Saveski et al. (2021) found that patterns of participant interaction systematically differ between conversations that eventually derail and ones that do not.

Accounting for social factors is important not only from the standpoint of practicality and improving model accuracy, but also from an AI ethics and social justice perspective: toxicity is highly dependent on social context, and phenomena such as microaggressions (Breitfeller et al., 2019) and dogwhistles (Mendelsohn et al., 2023) can alienate members of marginalized groups despite not being overtly toxic. Algorithmic tools for moderation and community governance that do not account for these social power dynamics (Sap et al., 2020) may therefore systematically fail to identify toxicity and other antisocial behavior directed towards marginalized groups, thereby exacerbating existing inequities (Davidson et al., 2017; Sap et al., 2019).

We know from prior work that capturing such social knowledge is at least feasible for algorithmic methods, with models trained specifically to predict phenomena such as patterns of participant interaction (Backstrom et al., 2013) and implicit hate speech (ElSherief et al., 2021) demonstrating good performance on those individual tasks. Our long-term challenge, then, is to find a way to incorporate all of this social knowledge, together with the linguistic factors we have already looked at, into a single forecasting algorithm. One possibly helpful insight is that social knowledge could be thought of as an additional modality that exists alongside language: the observed role of social context in augmenting purely linguistic meaning feels similar to how images (a more traditional example of an additional modality) can seamlessly integrate with text to produce combined meaning.[1] In this sense, it is encouraging to see that work on applying traditionally multimodal techniques to the task of forecasting derailment has seen some early success (Li et al., 2022), and we believe it is worth

[1] A perhaps whimsical example of how this plays out in the conversational domain is image-based memes, which can be inserted into otherwise text-based conversations and still result in an overall coherent dialogue; for more on this topic, see Milner (2012).

exploring how similar methodology could be applied to combine linguistic and social knowledge in forecasting.

6.2.3 Long-term Impact at Scale

The vision we have laid out for the big-picture role of tools like ConvoWizard (Section 6.1) is one where such tools form part of a broader ecosystem of tools for community governance and moderation, which when used together could reduce toxicity and improve the quality of conversations on online platforms. However, our findings thus far are still preliminary and have focused only on comment-level effects on a small slice of users. While the majority of users in our study reported that they *believe* that larger-scale adoption of ConvoWizard would have a net improvement on the overall quality of conversations in their community, it remains an open question whether this is truly the case. Our ultimate goal in this line of research, then, is to answer this question; that is, to understand the long-term, large-scale impact of our computational methods for promoting healthier online interactions.

More specifically, we would like to proceed from comment-level effects to the following progressively higher-level effects:

1. **Conversation-level**: Just because a user attempts to reduce tension—for instance, through the use of strategies such as politeness, objective language, factual framing, and question asking (Section 2.4.6)—does not mean they will actually succeed in reducing tension. The final outcome ultimately depends on how the user's comment is received by other users in the conversation. While the relatively small scale of the data from our

initial user study precludes a conversation-level analysis, a larger-scale study could enable a broad, data-driven exploration of conversational outcomes after a ConvoWizard intervention. It is possible that such an analysis might show that private user-facing interventions like those used by ConvoWizard are ineffectual but that more public interventions, like a moderator leaving a comment, might be more effective—in which case it would be worth exploring how computational methods could replicate such public interventions, for instance in the form of a bot that leaves public replies in conversations it deems to be at risk of derailment.

2. **User-level**: Even if proactive interventions within a conversation level successfully reduce the likelihood of derailment and improve conversation-level outcomes, moderators and good-faith users might hope for something more: that users exposed to interventions actually take away some lessons about how to avoid derailment. After all, results from our user study suggest that ignorance may be one reason that well-intentioned users end up making comments that increase tension: they sometimes honestly do not realize how their comment might be received negatively, a finding that has been echoed in other recent work (Srinivasan et al., 2019). Some of these users reported that ConvoWizard helped them realize this in specific instances, but does this translate to generalizable knowledge that persists across the user's future conversations? And if so, does the user actually heed the lesson, such that over time they become better at avoiding derailment on their own as opposed to relying on algorithmic tools as a crutch?

3. **Community-level**: Prior work has shown that toxic behavior can be contagious: frequent toxic behavior in a community can build a *culture* of

toxicity in which such behavior comes to be regarded as normal (Section 2.2.2). We may (optimistically) hope that the inverse is also true: that a preponderance of comments that employ tension-reduction techniques, and conversations that overwhelmingly end in a civil and friendly manner, may foster a culture of prosocial behavior. If tools like ConvoWizard turn out to successfully improve user-level and conversation-level outcomes (alongside comment-level ones we have observed), it can lay the theoretical groundwork, but it remains to be seen whether these concrete behavioral changes lead to a persistent culture change in the community. Undoubtedly, this would be the hardest effect to observe, and would require extremely large-scale studies taking place on vast timescales (as culture cannot change overnight). Nonetheless, it presents an idealistic long-term goal for this work—and it feels appropriate that we should conclude this dissertation on such high hopes.

APPENDIX A

MODERATOR INTERVIEW QUESTIONS

This appendix shows the general outline we followed for all moderator interviews. Note that this only served as a general guide; as the interview process is semi-structured we let the conversation flow naturally, so the exact order and wording of questions varied in practice.

A.1 Topic 1: Current Discussion Moderation Practices

- Understanding comment removal practices:
  - Q: How do you select comments to inspect for incivility and community rule violations?
  - Q: Do you ever proactively monitor ongoing conversations that you consider to be at risk of derailing into uncivil behavior?
  - Q: Say that you have a potentially problematic comment. Please describe your typical process for determining whether or not this comment needs moderation action.
    - (Optional) Q: Can you think of a specific example of a comment you took action on, and describe the process of determining whether or not that comment needed moderation action?
- Understanding how moderators use context:
  - Q: When you are considering [moderating/removing] a comment, do you generally read earlier comments in the thread for context? If so, what are you looking for?

  - Q: Do you think a user's [post and comment/edit] history affects your decision on whether you remove one of their comments? Can you give examples of such historical factors?
  - Q: After you take some moderation action on a comment, would you ever look at earlier comments in the conversation to identify more rule-violating comments?

- Understanding how moderators use automated tools:
  - Q: What automated tools do you currently use for moderation, if any?
  - Q: If you use any automated tools, do you use them for
    * triaging comments for you to review
    * *and/or*
    * automatically removing content?
    * If so, how do you configure automod for your community?

- Moderators' motivation and how they view the role of moderator:
  - Q: Why did you become a [moderator/administrator] for [Reddit/Wikipedia]?
  - Q: As an administrator, why did you become involved in discussion moderation work?

- Miscellaneous:
  - Q: How much time do you spend moderating each day? Each week?
    * First ask about time spent as an administrator, then ask about moderation.

- Q: How satisfied are you with your current moderation practices? Do you see room for improvement?
- Q: When doing your job as a moderator, would you rather:
  * (a) only [remove/take moderation action] very flagrant rule violations, and potentially miss some rule-breaking comments, or
  * (b) [remove/take moderation action] all comments that could be rule violations, potentially [remove/take moderation action] some comments that don't deserve to be.

A.2 Topic 2: Potential Use of Conversational Forecasting

- Understanding what moderators would do without time constraints:
  - Q: If you had more time for your job as a moderator, what actions would you want to do?
  - Q: If you had more time for your job as a moderator, when you make a moderation decision about a comment, would you read more of the context around the comment to inform your decision?
- Can moderators tell if a conversation is going awry? Can anyone?
  - Explanation: Here is some terminology that we will use for the rest of the interview:
    * We'll say that a comment is *civil* if it follows all the rules of your community, and that it is *uncivil* if it violates a community rule.
    * We will also say that a conversation *eventually derails* if it is civil right now, but in the future an uncivil comment gets posted to

the conversation.

  - Q: Given a civil conversation, do you think it is possible to foretell if a conversation will *eventually derail* into uncivil comments?
  - Q: Do you think you yourself are able to do this prediction?
    * If yes:
      · Q: Roughly how often do you think your prediction would be correct? That is, can you estimate what portion of the conversations you think will derail actually do end up derailing?
      · Q: What clues from a conversation do you use to inform your prediction?
  - Q: Do you think other moderators would be able to do this type of prediction?
  - Q: Do you think an algorithm might be able to do this type of prediction?
    * Q: Do you think it would be better or worse than humans?

- Monitoring derailing conversations:
  - Q: Assume you would know *for sure* that an ongoing conversation will turn uncivil in the future. Would you like to monitor new comments that are posted in this conversation?
  - Q: Now consider a more realistic scenario, where you cannot know for sure what the future of a conversation will be. Now, say we have a conversation that is predicted to derail; we will go through various levels of confidence in this prediction, and I want you to tell me if you would want to monitor new comments in the conversation for each level of prediction confidence.

    * Would you want to monitor new comments if you had *low* certainty in the prediction (i.e., *20%* of the conversations that are predicted to derail will eventually actually end up derailing)?
    * Would you want to monitor new comments if you had *50-50* certainty in the prediction (i.e., *50%* of the conversations that are predicted to derail will eventually actually end up derailing)?
    * Would you want to monitor new comments if you had *high* certainty in the prediction (i.e., *80%* of the conversations that are predicted to derail will eventually actually end up derailing)?

- Taking proactive steps for derailing conversations
  - Q: Assuming you would know for sure that a (currently civil) conversation will turn uncivil and violate the rules of the community, what proactive steps do you, as a moderator, see yourself taking in order to prevent uncivil behavior (if any)?
  - Q: Now—as before—consider a more realistic scenario, where you cannot know for sure what the future of a conversation will be. Now, say we have a conversation that is predicted to derail; we will go through various levels of confidence in this prediction, and I want you to tell me which of the proactive steps you just mentioned you would still take for each level of prediction confidence.
    * What proactive steps would you take if you had *low* certainty in the prediction (i.e., *20%* of the conversations that are predicted to derail will eventually actually end up derailing)?
    * What proactive steps would you take if you had *50-50* certainty in the prediction (i.e., *50%* of the conversations that are predicted

to derail will eventually actually end up derailing)?

    * What proactive steps would you take if you had *high* certainty in the prediction (i.e., *80%* of the conversations that are predicted to derail will eventually actually end up derailing)?

  - Q: Have you ever taken any of these proactive steps in the past?

A.3 Topic 3: Analyzing a Mockup Conversation

[The participant is shown a conversation from their community in the Conversation View, with the CRAFT score annotations removed.]

- Q: Do you think any comments in this conversation are uncivil?
- Q: How likely do you think this conversation is to eventually derail into uncivil behavior (breaking the rules of the community)? What made you think this way (point to specific behaviors)?
- Q: Would you want to monitor new comments in this conversation?
- Q: Would you consider taking any proactive steps to prevent uncivil behavior?

[The participant is shown the CRAFT scores for this conversation.]

- Ask the same questions again.

A.4 Topic 4: Analyzing a Mockup Ranking

[The moderator is shown a ranking of conversations from their community in the Ranking View.]

- Q: Which conversations do you think would be worth monitoring for uncivil behavior?
- Q: On the main page for the listing, how relevant is the information displayed about each thread?
- Q: What other information would you find useful in deciding whether to inspect or monitor a conversation?

APPENDIX B

CHANGEMYVIEW USER STUDY AND SURVEY DETAILS

B.1 Participant Recruitment

Participants for the study were recruited through two channels. First, the pinned announcement on ChangeMyView contained links for interested users to sign up for the study. Second, we direct messaged active members of ChangeMyView. Regardless of recruitment channel, all potential participants underwent a basic check of prior activity on ChangeMyView to filter out possible sockpuppet or brigader accounts, and also had to fill out a basic eligibility check to make sure that their typical ChangeMyView usage was compatible with ConvoWizard's technical limitations.[1] As an incentive for participation, $20 Amazon gift cards were offered to all participants who completed Phase 1 of the study, including filling out the exit survey. Across all participants who completed Phase 1, the mean *community age* (i.e., how long they had been active on ChangeMyView by the time of the study) was 3 years; the minimum was 3 months and the maximum was 8 years.

After Phase 1 was completed, we direct messaged all participants who had indicated in the exit survey that they would be interested in a follow-up study, inviting them to participate in Phase 2. For Phase 2, participants were given the option of participating for either a 30-day period (for which a $30 gift card incentive was offered) or a 60-day period (for which a $70 gift card incentive

[1] In particular, ConvoWizard's DOM-manipulation code was specifically engineered around the HTML structure of Reddit's classic desktop interface ("Old Reddit") and only works there, so users who primarily use other platforms (e.g., mobile) to access ChangeMyView would be ineligible.

was offered). All participants who accepted the invitation to join Phase 2 chose the 60-day option.

B.2 Exit Survey Implementation

The exit survey was implemented as a Qualtrics form, mostly consisting of multiple-choice questions with some optional free-response areas for participants to elaborate on their answers. The ConvoWizard tool automatically served the survey link to participants at the end of the 30-day period and participants could fill it out at any time after that, though we did send reminders via Reddit direct message.

B.3 Exit Survey Full Text and Raw Response Counts

Total Responses: 47

ConvoWizard Exit Survey

Thank you for your participation in the ConvoWizard study! As the final step in the study, we will now ask you a series of questions regarding your experience with ConvoWizard. The survey consists of a mix of multiple choice and free response questions. For free response questions, please provide as much information as you can. Your insights are extremely valuable in helping us with our research and, ultimately, with improving ConvoWizard.

After you submit this survey, we will follow up with your reward for partic-

ipation (a $20 Amazon gift card) via DM to the Reddit account you used to sign up for this study.

Q1: To begin, please enter your Reddit username. *[Free response]*

Part 2: Experiences with incivility on r/changemyview

The following questions will ask about your experiences with uncivil behavior on r/changemyview. For the purposes of this survey, "uncivil behavior" can be understood as comments that you judge to be violations of r/changemyview's Rule 2,* regardless of whether they ended up getting removed by moderators.
*Rule 2 says "Don't be rude or hostile to other users. Your comment will be removed even if the rest of it is solid. 'They started it' is not an excuse. You should report, not retaliate."

Q2: How big of a problem do you think incivility is on r/changemyview?

- It is almost nonexistent.: *3*
- It is only a minor problem.: *10*
- It is noticeable but not too big a problem.: *26*
- It is a pretty big problem.: *5*
- It is one of the biggest problems on the subreddit.: *3*

Q3: In your experience, what *most commonly* happens to uncivil comments on r/changemyview?

- I don't know (I have never seen any uncivil comments).: *2*

- They are removed by moderators.: *32*
- They are removed by the author.: *1*
- Nothing happens (the comment stays up).: *12*

Q4: In your experience, how quickly do r/changemyview moderators take action on uncivil comments?

- I have never seen moderators take action on uncivil comments.: *3*
- They act almost immediately after the comment is posted.: *4*
- They act within a few hours after the comment is posted.: *25*
- They act within the day the comment is posted (but take more than a few hours).: *13*
- They take more than a day to act.: *2*

Q5: In your experience, what *most commonly* happens to discussions on r/changemyview after an uncivil comment gets posted and is not immediately removed?

- I don't know (I have never seen any uncivil comments, or every uncivil comment I've seen was immediately removed).: *2*
- The situation escalates and more uncivil replies are posted.: *22*
- The situation recovers and becomes civil again.: *5*
- The discussion dies and no further replies are posted.: *18*

Show the following question(s) if "The situation escalates and more uncivil replies are posted" was selected in ***Q5*** *(22 participants):*

Q6: In discussions that you've seen escalate after an uncivil comment was posted and not immediately removed, what *most commonly* happens if the comment is eventually removed?

- I don't know (I have never seen an uncivil comment get removed).: *1*
- The removal helps the situation to recover.: *4*
- The removal has no effect because it is ignored by the people in the discussion.: *7*
- The removal has no effect because the discussion has already ended.: *10*

Q7: Have you ever made a comment on r/changemyview that you later regretted because in hindsight it could be perceived as offensive or uncivil?

- Never.: *15*
- Yes, and the moderators removed it.: *4*
- Yes, and I later removed it myself.: *19*
- Yes, and it was never removed.: *9*

Q8: Which of the following statements about r/changemyview's enforcement of Rule 2 do you agree with? (Check all that apply)

- I am satisfied with the existing enforcement.: *28*
- The existing enforcement is too much (comments often get removed that didn't deserve it).: *7*
- The existing enforcement is not enough (comments that deserve to be removed often aren't).: *9*

- The existing enforcement is biased.: *6*
- It is too hard to get a bad enforcement decision overturned.: *4*
- Enforcement needs to be more transparent.: *16*

Q9: Are there any other things you wish r/changemyview did differently in enforcing Rule 2? *[Free response (See Section B.4 for sampled answers)]*

Part 3: Forecasting incivility

The following questions will ask about your personal intuitions about when incivility occurs in discussions. We emphasize that you should answer these questions from the perspective of your own intuitions, **without** the help of ConvoWizard.

Q10: Can you personally tell when discussions are at risk of turning uncivil (that is, may later lead to comments that will violate Rule 2)?

- I cannot tell.: *0*
- I can tell in some cases.: *19*
- I can tell in many cases.: *18*
- I can tell in most cases.: *10*

*Show the following question(s) if "I cannot tell" was NOT selected in **Q10** (47 participants):*

Q11: Briefly explain how you can tell if a discussion is at risk of turning uncivil. *[Free response]*

Q12: If you think a discussion is at risk of turning uncivil, does this make you more willing or less willing to participate?

- More willing: *2*
- Less willing: *29*
- No effect: *16*

Q13: If you think a discussion is at risk of turning uncivil and you are participating, does this affect how you phrase your comments?

- Yes: *36*
- No: *11*

*Show the following question(s) if "Yes" was selected in **Q13** (36 participants):*

Q14: How does the phrasing you use in your comments change when you think the discussion is at risk of turning uncivil? Select all that apply:

- I use more polite language.: *19*
- I use fewer swear words.: *2*
- I use more formal language.: *17*
- I use more casual language.: *4*
- I use more objective language (that is, I try to frame my comment in terms of facts and data).: *24*
- I use more subjective language (that is, I try to frame my comment in terms of personal feelings and opinions).: *4*
- I ask more questions.: *18*
- I write a shorter comment.: *10*
- I write a longer comment.: *11*

- Other (please describe):: *9*

Part 4: Experience with ConvoWizard: Context Summary Feedback

The following questions will ask about your experience with the Context summary feedback feature of ConvoWizard. This is referring to the top box that gave a summary of how likely the preexisting discussion was to turn uncivil before you joined (see the highlighted part of the screenshot below): *[Screenshot of ConvoWizard interface with Context Summary box highlighted]*

Q15: Do you remember seeing the text and/or color of the context summary box change (indicating that the discussion might be getting tense)?

- Yes: *38*
- No: *9*

Show the following question(s) if "Yes" was selected in ***Q15*** *(38 participants):*

Q16: Thinking specifically of times when you saw the text and/or color of the context summary box change, did the context summary feedback ever...

a) ...help you avoid a fight or confrontation?

- Yes: *19*
- No: *19*

b) ...affect whether you decided to post a reply?

- Yes: *20*
- No: *18*

c) ...affect what you said in your reply, if you posted one?

- Yes: *26*
- No: *12*

Show the following question(s) if "Yes" was selected in ***Q16c*** *(26 participants):*

Q17: Thinking specifically of times when you saw the text and/or color of the context summary box change, how did the context summary feedback affect what you said in your reply? Select all that apply:

- I used more polite language.: *17*
- I used fewer swear words.: *1*
- I used more formal language.: *7*
- I used more casual language.: *3*
- I used more objective language (that is, I try to frame my comment in terms of facts and data).: *9*
- I used more subjective language (that is, I try to frame my comment in terms of personal feelings and experiences).: *2*
- I asked more questions.: *9*
- I wrote a shorter comment.: *8*
- I wrote a longer comment.: *2*
- Other (please describe): *4*

Q18: Overall, how useful was the context summary feedback?

- Not at all useful: *8*

- Somewhat useful: *19*
- Quite useful: *10*
- Very useful: *1*

Q19: Do you think ConvoWizard is better or worse than you at telling whether a discussion might be getting tense?

- Much better: *2*
- Somewhat better: *7*
- About the same: *16*
- Somewhat worse: *15*
- Much worse: *7*

*Show the following question(s) if "Much better" or "Somewhat better" was selected in **Q19** (9 participants):*

Q20: Why do you think ConvoWizard is better than you at telling whether a discussion might be getting tense? *[Free response]*

*Show the following question(s) if "Much worse" or "Somewhat worse" was selected in **Q19** (22 participants):*

Q21: Why do you think ConvoWizard is worse than you at telling whether a discussion might be getting tense? *[Free response]*

Q22: For which of the following reasons, if any, did you ever disagree with the context summary feedback? "Disagree" means that you intuitively felt the feedback was wrong, or you would have made a different judgment

call. Rate how often each potential disagreement occurred on a scale from "Never" to "Very often".

a) ConvoWizard said a discussion looked tense even though it wasn't
    - Never: *6*
    - Rarely: *12*
    - Sometimes: *21*
    - Often: *7*
    - Very often: *1*

b) ConvoWizard did not say a discussion was tense even though it clearly was.
    - Never: *14*
    - Rarely: *18*
    - Sometimes: *14*
    - Often: *0*
    - Very often: *1*

c) ConvoWizard's estimated degree of tension was incorrect (for example, a discussion was marked as "somewhat" tense when it was actually extremely tense).
    - Never: *14*
    - Rarely: *13*
    - Sometimes: *13*
    - Often: *5*
    - Very often: *2*

d) ConvoWizard's context summary feedback seemed to be biased.

- Never: *27*
- Rarely: *13*
- Sometimes: *6*
- Often: *1*
- Very often: *0*

Q23: Are there any other reasons not listed above that you disagreed with the context summary feedback? (You can also use this space to elaborate on your answers to the previous question). *[Free response]*

Part 5: Experience with ConvoWizard: Reply Summary Feedback

The following questions will ask about your experience with the Reply summary feedback feature of ConvoWizard. This is referring to the bottom box that gave a summary of how the reply you were drafting could affect the tension in the discussion if it was posted (see the highlighted part of the screenshot below): *[Screenshot of ConvoWizard with the Reply Summary box highlighted]*

Q24: Do you remember seeing the text and/or color of the reply summary box change (indicating potential increase or decrease in tension)?

- Yes: *35*
- No: *12*

Show the following question(s) if "Yes" was selected in ***Q24*** *(35 participants):*

Q25: Thinking specifically of times when you saw the text and/or color of the reply summary box change, did the reply summary

feedback ever...

a) ...help you avoid a fight or confrontation?
   - Yes: *19*
   - No: *16*

b) ...stop you from posting something you might have regretted later?
   - Yes: *19*
   - No: *16*

c) ...affect whether you decided to eventually post your draft reply?
   - Yes: *21*
   - No: *14*

d) ...affect what you said in the reply you ended up posting, if you posted one?
   - Yes: *25*
   - No: *10*

Show the following question(s) if "Yes" was selected in ***Q25c*** *(25 participants):*

Q26: Thinking specifically of times when you saw the text and/or color of the reply summary box change to indicate an increase in tension (i.e. a reddish color), how did the reply summary feedback change what you said in your reply? Select all that apply:
   - N/A (I have never seen an increase in tension).: *0*

- I used more polite language.: *17*
- I used fewer swear words.: *2*
- I used more formal language.: *12*
- I used more casual language.: *5*
- I used more objective language (that is, I try to frame my comment in terms of facts and data).: *11*
- I used more subjective language (that is, I try to frame my comment in terms of personal feelings and experiences).: *1*
- I asked more questions.: *8*
- I wrote a shorter comment.: *4*
- I wrote a longer comment.: *4*
- Other (please describe): *3*

Q27: Overall, how useful was the reply summary feedback?

- Not at all useful: *8*
- Somewhat useful: *17*
- Quite useful: *10*
- Very useful: *0*

Q28: Do you think ConvoWizard is better or worse than you at telling whether a draft reply might increase tension in the discussion?

- Much better: *1*
- Somewhat better: *8*
- About the same: *23*

- Somewhat worse: *12*
- Much worse: *3*

Show the following question(s) if "Much better" or "Somewhat better" was selected in ***Q28*** *(9 participants):*

Q29: Why do you think ConvoWizard is better than you at telling whether a draft reply might increase tension in the discussion? *[Free response]*

Show the following question(s) if "Much worse" or "Somewhat worse" was selected in ***Q28*** *(15 participants):*

Q30: Why do you think ConvoWizard is worse than you at telling whether a draft reply might increase tension in the discussion? *[Free response]*

Q31: For which of the following reasons, if any, did you ever disagree with the reply summary feedback? "Disagree" means that you intuitively felt the feedback was wrong, or you would have made a different judgment call. Rate how often each potential disagreement occurred on a scale from "Never" to "Very often".

a) ConvoWizard said my reply would increase tension even though it clearly wouldn't.
   - Never: *9*
   - Rarely: *9*
   - Sometimes: *20*

- Often: *6*
- Very often: *3*

b) ConvoWizard did not say my reply would increase tension even though it clearly would.

- Never: *20*
- Rarely: *11*
- Sometimes: *13*
- Often: *2*
- Very often: *1*

c) Changing the text of my draft did not seem to change what ConvoWizard said.

- Never: *13*
- Rarely: *12*
- Sometimes: *15*
- Often: *6*
- Very often: *1*

d) A minor/trivial change to the text of my draft changed what ConvoWizard said.

- Never: *11*
- Rarely: *7*
- Sometimes: *15*
- Often: *9*
- Very often: *5*

e) ConvoWizard's reply summary feedback seemed to be biased.

- Never: *27*
- Rarely: *12*
- Sometimes: *8*
- Often: *0*
- Very often: *0*

Q32: Are there any other reasons not listed above that you disagreed with the reply summary feedback? (You can also use this space to elaborate on your answers to the previous question). *[Free response]*

Part 6: Overall impressions

The following questions ask about your overall impressions of ConvoWizard, accounting for all its features.

Q33: Between the context summary feedback and reply summary feedback, which did you find more helpful?

- Context summary: *7*
- Reply summary: *18*
- Both were equally helpful: *8*
- Both were equally unhelpful: *14*

Q34: If ConvoWizard were to be publicly released and worked on all versions of Reddit (including new Reddit and mobile), how likely would you be to use it as part of your usual r/changemyiew participation?

- I would definitely not use it.: *8*
- I might try it.: *18*

- I would probably try it.: *13*
- I would definitely use it.: *3*
- I would definitely use it, and recommend it to others.: *5*

Q35: If ConvoWizard were to be publicly released and many members of r/changemyview used it, do you think this would improve or harm overall discussion quality?

- It would improve discussion quality: *30*
- It would harm discussion quality: *1*
- It would have little to no effect: *16*

Q36: Which would you prefer to use: ConvoWizard (which predicts whether a discussion / comment might lead to uncivil behavior in the future), or a tool that detects whether a discussion/comment is already uncivil?

- I would prefer ConvoWizard.: *25*
- I would prefer the tool that detects already existing incivility.: *5*
- I would use both.: *7*
- I would use neither.: *10*
- I cannot tell the difference.: *0*

Q37: Which of the following improvements would be most important to you in deciding to use or recommend ConvoWizard? (Select up to 3)

- Correctly identifying more of the tense discussions or draft replies.: *18*
- Giving fewer false alerts on harmless discussions or replies.: *17*

- Better user interface and integration with the Reddit webpage.: *14*
- More consistent behavior.: *7*
- More transparency (i.e., explanations of why ConvoWizard marked a discussion / comment as tense).: *23*
- More concrete suggestions on how to decrease tension: *14*
- Availability on other platforms (new Reddit, mobile app, etc.).: *13*
- Other (please describe): *6*

Q38: Did you encounter any technical issues while using ConvoWizard?

- Yes (please describe): *0*
- No: *37*

Q39: Would you be interested in continuing to test ConvoWizard, assuming we extend the testing period? This is entirely optional and the answer to this question will not affect your receipt of the $20 gift card for the testing period you just finished.

- Yes: *33*
- No: *14*

Q40: In the case the results of this study will be published in a scientific article, would you be OK with us anonymously quoting your answers you provided in this survey? We will not disclose your Reddit username (or any other identity).

- Yes: *44*
- No: *3*

B.4 Sampled Free Responses

For each free response question, we have randomly sampled three responses to be shown as examples (unless there were fewer than three total responses, for optional / conditional questions).

Are there any other things you wish r/changemyview did differently in enforcing Rule 2?

- Being consistent. CMV removes certain comments, but far after the conversation dissolves into insults and hostility.
- There are clearly a large bias present in the subreddit, particularly on topics that if you are not going along with what is the 'popular' thing then you get downvoted, or just insulted.
- No, the moderators are great with enforcement.

Briefly explain how you can tell if a discussion is at risk of turning uncivil.

- Just a feeling that some people are starting more hostile than others.
- If the conversation starts getting personal, attacking personal credentials or identity instead of the problem.
- The easiest way is to analyze the phrasing. Stern, short phrases, completely contradicting the other person's viewpoint might come off as hostile and aggressive, causing a defensive reaction that might turn into an uncivil discussion.

How does the phrasing you use in your comments change when you think the discussion is at risk of turning uncivil? Select all that apply: - Other (please describe):

- I give minor concessions to points they have made
- try to explain why you see the way you do and what makes you disagree with them.
- Pretty much all of the above to some degree. I never want to offend anyone. And I try not to be offended. Swearing just turns things instantly uncivil.

Thinking specifically of times when you saw the text and/or color of the context summary box change, how did the context summary feedback affect what you said in your reply? Select all that apply: - Other (please describe)

- All of the above again. I thought of better words I could usem maybe words that don't sound like I may be trying to provoke a uncivil response. I tried longer comments as I am not good at summarizing things in short comments. I enjoyed having this tool to help me see things I may not realize I am posting.
- I kept rewording my reply until it stopped showing up orange. It usually led to less effective replies that, in retrospect, were too wishy-washy to change anyone's view.
- I tended to avoid certain key words that I felt the program picked up on whether or not I was being confrontational. The word "you" or any words with negative connotations could be altered without changing the meat of my messages.

Why do you think ConvoWizard is better than you at telling whether a discussion might be getting tense?

- Often times if the color changed I would reread what I was saying and see if the response maybe cam off the wrong way. Helping me then to reword it.
- It seems to be able to sense strong emotions, but it doesn't seem to understand pathos arguments.
- It's hard in the moment when reading a divisive comment to objectively recognize where the conversation is going

Why do you think ConvoWizard is worse than you at telling whether a discussion might be getting tense?

- I tried to test its capabilities. In my experience, direct insults do not necessarily alert the program of anything being wrong. The comment has to be sufficiently long for it to usually detect possible cases of uncivility rising. It also seems a little too sensitive, sometimes a comment that was meant to be stern alerts ConvoWizard.
- ConvoWizard seemed to be based off of specific words being in the conversation at all? Discussion on the r-slur were always red, because the word set ConvoWizard off. Quoting other people's tens dialog also seemed to affect that Wizard just as much as saying it myself, but quoting other people's dialogue is just required to have the discussion.
- It said everything was at risk of getting tense

Are there any other reasons not listed above that you disagreed with the context summary feedback? (You can also use this space to elaborate on your answers to the previous question).

- Yes, sometimes the conversation was becoming tense and ConvoWizard didn't notice it.
- Frankly, I just didn't encounter many tense arguments. I was impressed with the tool's sentiment analysis, but I don't have any evidence that it could identify tension that I or most other commenters would fail to identify.
- It got obvious things right but didn't seem to work well on the fringe cases.

Thinking specifically of times when you saw the text and/or color of the reply summary box change to indicate an increase in tension (i.e. a reddish color), how did the reply summary feedback change what you said in your reply? Select all that apply: - Other (please describe)

- See previous answer. I reworded it. Looking back, I disagree with my rewording and think my posts became less likely to earn a delta.
- I'm not entirely sure what changed but that I did

Why do you think ConvoWizard is better than you at telling whether a draft reply might increase tension in the discussion?

- It's easy to pick up the read tense but I'm not always sure when what I'm going to say will make things better or worse.

- Certain verbiage that I typically used the wizard pointed out and I adjusted the verbiage.
- I don't often care about increasing tension. My objective is generally the discussion, not whether I sound polite or not. ConvoWizard sort of reminds me that I should use maybe different language.

Why do you think ConvoWizard is worse than you at telling whether a draft reply might increase tension in the discussion?

- Sometimes it seemed to think a very innocuous response would escalate tension when I found that unlikely.
- I just don't think the extension works very well. It must be a technical issue.
- It reacted to obvious stimuli but didn't work well with sarcasm or curtness, which are often the first signs that a conversation is becoming tense.

Are there any other reasons not listed above that you disagreed with the reply summary feedback? (You can also use this space to elaborate on your answers to the previous question).

- No.
- Just as before, when quoting someone else's text, ConvoWizard treated it as if the person themselves was saying it. This misrepresents the discussion.
- Primarily that it seemed to say everything was in danger of tension

Which of the following improvements would be most important to you in deciding to use or recommend ConvoWizard? (Select up to 3) - Other (please describe)

- Firefox please.
- I'm perfectly able to tell if people are getting 'tense'. I don't need software to tell me.
- Honestly, the biggest issue is me. I almost always knew when a conversation was getting uncivil, but was going to post regardless. The wizard rarely shamed me into not posting (although it did work occasionally! which was surprising). Granted, i do use reddit as an outlet to vent/argue, so i wasn't really trying to avoid being uncivil. If it doesn't change my behavior, it doesn't do much to warn me something is uncivil

Did you encounter any technical issues while using ConvoWizard? - Yes (please describe)

- It occasionally would stop returning a result mid-reply, or not really return a result at all.
- My anti-virus flagged it once.
- text boxes that are light gray on white. v hard to read.

Is there any additional feedback you would like to provide that was not already covered, or anything in particular that you liked or disliked?

- I would love to see this as a feature on Reddit in general. It could really help things. Though if it would change how people act is unseen.

- It was hard to use, since I had to use old Reddit. It made me use it less often
- no

APPENDIX C

DETAILS ON DERAILMENT ANNOTATION PROCEDURE

The process of constructing a labeled dataset for personal attacks was challenging due to the complex and subjective nature of the phenomenon, and developed over several iterations as a result. In order to guide future work, here we provide a detailed explanation of this process, expanding on the description in Section 3.3.

Our goal was to understand linguistic markers of conversations that derail into personal attacks—a highly subjective phenomenon with a multitude of possible definitions.[1] To enable a concrete analysis of conversational derailment that encompasses the scale and diversity of a setting like Wikipedia talk pages, we therefore needed to develop a well-defined conceptualization of derailment, and a procedure to accurately discover instances of this phenomenon at scale.

Our approach started from an initial qualitative investigation that resulted in a seed set of example derailments. This seed set then informed the design of the subsequent crowdsourced filtering procedure, which we used to construct our full dataset.

C.1 Initial qualitative investigation

To develop our task, we compiled an initial sample of potentially derailing conversations by applying the candidate selection procedure (detailed in Section 3.3) to a random subset of Wikipedia talk pages. This procedure yielded a set

[1] Refer to Turnbull (2018) for examples of challenges community moderators face in delineating personal attacks.

of conversations which the underlying trained classifier deemed to be initially civil, but with a later toxic comment. An informal inspection of these candidate conversations suggested many possible forms of toxic behavior, ranging from personal attacks ('Are you that big of a coward?'), to uncivil disagreements ('Read the previous discussions before bringing up this stupid suggestion again.'), to generalized attacks ('Another left wing inquisition?') and even to outright vandalism ('Wikipedia SUCKS!') or simply unnecessary use of foul language.

Through our manual inspection, we also identified a few salient points of divergence between the classifier and our (human) judgment of toxicity. In particular, several comments which were machine-labeled as toxic were clearly sarcastic or self-deprecating, perhaps employing seemingly aggressive or foul language to bolster the collegial nature of the interaction rather than to undermine it. These false positive instances highlight the necessity of the subsequent crowdsourced vetting process—and point to opportunities to enrich the subtle linguistic and interactional cues such classifiers can address.

Seed set. Our initial exploration of the automatically discovered candidate conversations and our discussions with the members of the Wikimedia Foundation anti-harassment program pointed to a particularly salient and perplexing form of toxic behavior around which we centered our subsequent investigation: personal attacks *from within*, where one of the two participants of the ostensibly civil initial exchange turns on another interlocutor. For each conversation where the author of the toxic-labeled comment also wrote the first or second comment, the authors manually checked that the interaction started civil and ended in a personal attack. The combined automatic and manual filtering process resulted in

our seed set of 232 derailing conversations.

We additionally used the candidate selection procedure to obtain on-track counterparts to each conversation in the seed set that took place on the same talk-page; this pairing protocol is further detailed in Section 3.3.

Human performance. We gaged the feasibility of our task of predicting future personal attacks by asking (non-author) volunteer human annotators to label a 100-pair subset of the seed set. In this informal setting, also described in Section 3.6, we asked each annotator to guess which conversation in a pair will lead to a personal attack on the basis of the initial exchange. Taking the majority vote across three annotators, the human guesses achieved an accuracy of 72%, demonstrating that humans indeed have some systematic intuition for a conversation's potential for derailment.

Informing the crowdsourcing procedure. To scale beyond the initial sample, we sought to use crowdworkers to replicate our process of manually filtering automatically-discovered candidates, enabling us to vet machine-labeled awry-turning and on-track conversations across the entire dataset. Starting from our seed set, we adopted an iterative approach to formulate our crowdsourcing tasks.

In particular, we designed an initial set of task instructions—along with definitions and examples of personal attacks—based on our observations of the seed set. Additionally, we chose a subset of conversations from the seed set to use as *test questions* that crowdworker judgements on the presence or absence of such behaviors could be compared against. These test questions served both as anchors to ensure the clarity of our instructions, and as quality controls. Mis-

matches between crowdworker responses and our own labels in trial runs then motivated subsequent modifications we made to the task design. The crowdsourcing jobs we ultimately used to compile our entire dataset are detailed below.

C.2 Crowdsourced filtering

Based on our experiences in constructing and examining the seed set, we designed a crowdsourcing procedure to construct a larger set of personal attacks. Here we provide more details about the crowdsourcing tasks, outlined in Section 3.3. We split the crowdsourcing procedure into two jobs, mirroring the manual process used to construct the seed set outlined above. The first job selected conversations ending with personal attacks; the second job enforced that derailing conversations start civil, and that on-track conversations remain civil throughout. We used the CrowdFlower platform[2] to implement and deploy these jobs.

Job 1: Ends in personal attack. The first crowdsourcing job was designed to select conversations containing a personal attack. In the annotation interface, each of three annotators was shown a candidate derailing conversation (selected using the procedure described in Section 3.3). The suspected toxic comment was highlighted, and workers were asked whether the highlighted comment contains a personal attack—defined in the instructions as a comment that is "rude, insulting, or disrespectful towards a person/group or towards that person/group's actions, comments, or work." We instructed the annotators not

[2]This platform is now defunct.

to confuse personal attacks with civil disagreement, providing examples that illustrated this distinction.

To control the quality of the annotators and their responses, we selected 82 conversations from the seed set to use as *test questions* with a known label. Half of these test questions contained a personal attack and the other half were known to be civil. The CrowdFlower platform's quality control tools automatically blocked workers who missed at least 20% of these test questions.

While our task sought to identify personal attacks towards other interlocutors, trial runs of Job 1 suggested that many annotators construed attacks directed at other targets—such as groups or the Wikipedia platform in general—as personal attacks as well. To clarify the distinction between attack targets, and focus the annotators on labeling personal attacks, we asked annotators to specify *who* the target of the attack is: (a) someone else in the conversation, (b) someone outside the conversation, (c) a group, or (d) other. The resultant responses allowed us to filter annotations based on the reported target. This question also played the secondary role of ensuring that annotators read the entire conversation and accounted for this additional context in their choice.

In order to calibrate annotator judgements of what constituted an attack, we enforced that annotators saw a reasonable balance of awry-turning and on-track conversations. By virtue of the candidate selection procedure, a large proportion of the conversations in the candidate set contained attacks. Hence, we also included 804 candidate on-track conversations in the task.

Using the output of Job 1, we filtered our candidate set to the conversations where *all three annotations* agreed that a personal attack had occurred. We found

that unanimity produced higher quality labels than taking a majority vote by omitting ambiguous cases (e.g., the comment "It's our job to document things that have received attention, however ridiculous we find them." could be insulting towards the things being documented, but could also be read as a statement of policy).[3]

Job 2: Civil start. The second crowdsourcing job was designed to enforce that candidate derailing conversations start civil, and candidate on-track conversations remain civil throughout. Each of three annotators was shown comments from both on-track and derailing conversations that had already been filtered through Job 1. They were asked whether any of the displayed comments were toxic—defined as "a rude, insulting, or disrespectful comment that is likely to make someone leave a discussion, engage in fights, or give up on sharing their perspective." This definition was adapted from previous efforts to annotate toxic behavior Wulczyn et al. (2017) and intentionally targets a broader spectrum of uncivil behavior.

As in Job 1, we instructed annotators to not confound civil disagreement with toxicity. To reinforce this distinction, we included an additional question asking them whether any of the comments displayed disagreement, and prompted them to identify particular comments.

Since toxicity can be context-dependent, we wanted annotators to have access to the full conversation to help inform their judgement about each comment. However, we were also concerned that annotators would be overwhelmed by the amount of text in long conversations, and might be deterred from carefully reading each comment as a result. Indeed, in a trial run where

[3]This choice further sacrifices recall for the sake of label precision, an issue that is also discussed in Section 3.7.

full conversations were shown, we received negative feedback from annotators regarding task difficulty. To mitigate this difficulty without entirely omitting contextual information, we divided each conversation into snippets of three comments each. This kept the task fairly readable while still providing some local context. For candidate derailing conversations, we generated the snippets from all comments except the last one (which is known from Job 1 to be an attack). For on-track conversations, we generated the snippets from all comments in the conversation.

We marked conversations as toxic if at least three annotators, across all snippets of the conversation, identified at least one toxic comment. As in Job 1, we found that requiring this level of consensus among annotators produced reasonably high-quality labels.

Overall flow. To compile our full dataset, we started with 3,218 candidate derailing conversations which were filtered using Job 1, and discarded all but 435 conversations which all three annotators labeled as ending in a personal attack towards someone else in the conversation. These 435 conversations, along with paired on-track conversations, were then filtered using Job 2. This step removed 30 pairs: 24 where the derailing conversation was found to contain toxicity before the personal attack happened, and 6 where the on-track conversation was found to contain toxicity. We combined the crowdsourced output with the seed set, then did a final round of our own manual validation, to obtain a final dataset of 1,168 paired derailing and on-track conversations.

APPENDIX D

FURTHER EXAMPLES OF PROMPT TYPES

Table D.1 provides further examples of comments containing the prompt types we automatically extracted from talk page conversations using the unsupervised methodology described in Section 3.4; descriptions of each type can be found in Table 3.2. For additional interpretability, we also include examples of typical *replies* to comments of each prompt type, which are also extracted by the method.[1]

[1]Note that the comment and reply examples in each row of the table do not necessarily correspond to one another.

Prompt Type	Example comments	Example replies
`Factual check`	I do **assert** that you **are wrong**. The caption **states that** [...] This **is true** whether the tax system is progressive or regressive.	**I agree** with you on some points. **Please clarify**; I read the german, which is why I re-added the term.
`Moderation`	Whatever your reasons, **please stop reverting**, and take the issue to [link]. You have undone many redirects against consensus. **This is clearly disruptive** [...]	The correct merge procedure **was followed**. **I apologize** for insulting other users [...] I will **never do that** again.
`Coordination`	**I can** probably **help out** with this. I was wondering if you wanted to **work together** on another project.	**Great**! Thanks for your speedy response. **OK**. I've **just started** reviewing them.
`Casual remark`	I did **kinda** forget that. It's gone now **anyway**... Hey, thanks. And, **yeah,** I'm back. **I had** a very pleasant break.	**Lol** - you were quick tonight with your reverts. Sorry. It just looks **so cool**! **Oh well**. **Anyway**, thank for the heads up.
`Action statement`	**Do you know** when they will be **added**? I've also **found** a couple of sources that give his birth year as 1867, so that could probably **be changed**.	**Ok**. I **found** it, and **added** the link myself. **Great**, I **looked into** the code and **changed** the template.
`Procedures`	**Please** could you shed some light on why you **restored** it. You may **delete** the aforementioned image. I **have uploaded** a new one [...]	The image **is not copyrighted**, so I [...] **restored** it to your page. If these are not **added** within 24 hours I will **delete** it again.

Table D.1: Further examples of representative comments in the data for each automatically-extracted prompt type, and examples of typical replies prompted by each type, produced by the methodology in Section 3.4. Bolding indicates common phrasings identified by the framework in the respective examples.

APPENDIX E

CRAFT AND BERT VARIANCE STATISTICS

	(a) CGA-WIKI					(b) CGA-CMV				
Run	A	P	R	FPR	F1	A	P	R	FPR	F1
					CRAFT					
0	64.2	61.6	75.0	46.7	67.7	60.5	57.6	79.7	58.8	66.8
1	65.5	63.0	75.0	44.0	68.5	62.0	59.7	73.7	49.7	66.0
2	64.8	62.3	75.0	45.5	68.0	60.9	57.8	80.4	58.6	67.3
3	65.2	63.3	72.4	41.9	67.6	62.1	59.8	73.8	49.6	66.1
4	64.5	62.3	73.3	44.3	67.4	62.6	61.0	70.0	44.7	65.2
5	66.3	64.5	72.4	39.8	68.2	60.6	57.5	80.8	59.6	67.2
6	63.3	60.9	74.3	47.6	67.0	59.9	57.1	79.5	59.8	66.5
7	65.7	63.9	72.1	40.7	67.8	61.9	59.9	72.2	48.4	65.5
8	64.4	61.8	75.5	46.7	68.0	61.6	59.0	76.0	52.8	66.5
9	65.0	63.3	71.4	41.4	67.1	61.2	58.7	75.1	52.8	65.9
Mean	64.9	62.7	73.6	43.9	67.7	61.3	58.8	76.1	53.5	66.3
SD	0.9	1.1	1.5	2.8	0.5	0.9	1.3	3.8	5.4	0.7
					BERT					
0	62.3	60.7	69.5	45.0	64.8	62.8	61.5	68.6	43.0	64.8
1	61.9	60.0	71.7	47.9	65.3	61.4	59.5	71.3	48.5	64.9
2	63.7	61.1	75.5	48.1	67.5	62.1	59.7	74.1	50.0	66.1
3	61.3	59.7	69.5	46.9	64.2	62.1	61.6	64.2	40.1	62.8
4	65.5	63.3	73.8	42.9	68.1	61.1	59.0	72.7	50.4	65.1
5	58.2	57.1	66.0	49.5	61.2	61.4	60.8	64.0	41.2	62.4
6	66.2	64.7	71.4	39.0	67.9	61.3	59.5	70.8	48.2	64.6
7	64.4	63.1	69.5	40.7	66.1	60.6	58.6	72.1	50.9	64.7
8	61.5	62.5	57.6	34.5	60.0	61.5	60.0	69.0	46.1	64.2
9	63.8	62.8	67.6	40.0	65.1	62.6	60.2	74.4	49.1	66.6
Mean	62.9	61.5	69.2	43.5	65.0	61.7	60.0	70.1	46.8	64.6
SD	2.3	2.2	4.9	4.8	2.7	0.7	1.0	3.7	4.0	1.3

Table E.1: Variance of (A)ccuracy, (P)recision, (R)ecall, False Positive Rate (FPR), and F1 for 10 runs of CRAFT (top) and BERT (bottom) on the (a) CGA-WIKI and (b) CGA-CMV datasets. Mean and standard deviation across all 10 runs are also reported.

As described in Section 4.6.3, the nondeterminism present in neural network training methods leads to a small amount of variance in the results of training

CRAFT and the Sliding BERT baseline. To quantify this variance, we run 10 rounds of CRAFT and BERT fine-tuning from scratch on both the CGA-WIKI and CGA-CMV datasets. The measured performance metrics in all 10 runs, as well as the mean and standard deviation, are enumerated in Table E.1.

BIBLIOGRAPHY

Rob Abbott, Marilyn Walker, Pranav Anand, Jean E. Fox Tree, Robeson Bowmani, and Joseph King. How Can You Say Such Things?!?: Recognizing Disagreement in Informal Political Argument. In *Proceedings of the Workshop on Languages in Social Media*, 2011.

Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of CHI*, 2018.

Amina Adadi and Mohammed Berrada. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, 2018.

Yavuz Akbulut, Yusuf Levent Sahin, and Bahadir Eristi. Cyberbullying Victimization among Turkish Online Social Utility Members. *Educational Technology & Society*, 13(4), October 2010.

Muhammad Ali, Angelica Goetzen, Alan Mislove, Elissa M. Redmiles, and Piotr Sapiezynski. Problematic Advertising and its Disparate Exposure on Facebook. In *32nd USENIX Security Symposium (USENIX Security 23)*, 2023.

Jennifer Allen, Markus Mobius, David M. Rothschild, and Duncan J. Watts. Research note: Examining potential bias in large-scale censored data. *Harvard Kennedy School Misinformation Review*, July 2021.

Kelsey Allen, Giuseppe Carenini, and Raymond T. Ng. Detecting Disagreement in Conversations using Pseudo-Monologic Rhetorical Structure. In *Proceedings of EMNLP*, 2014.

Tim Althoff, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. How to Ask

for a Favor: A Case Study on the Success of Altruistic Requests. In *Proceedings of ICWSM*, 2014.

Tim Althoff, Kevin Clark, and Jure Leskovec. Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health. *Transactions of the Association for Computational Linguistics*, 4, 2016.

Ofer Arazy, Lisa Yeo, and Oded Nov. Stay on the Wikipedia Task: When Task-related Disagreements Slip Into Personal and Procedural Conflicts. *J. Assoc. Inf. Sci. Technol.*, 64(8), August 2013.

Lisa P. Argyle, Christopher A. Bail, Ethan C. Busby, Joshua R. Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences*, 120(41), October 2023.

Zahra Ashktorab and Jessica Vitak. Designing Cyberbullying Mitigation and Prevention Solutions through Participatory Design With Teenagers. In *Proceedings of CHI*, 2016.

Malika Aubakirova and Mohit Bansal. Interpreting Neural Networks to Improve Politeness Comprehension. In *Proceedings of EMNLP*, 2016.

Lars Backstrom, Jon Kleinberg, Lillian Lee, and Cristian Danescu-Niculescu-Mizil. Characterizing and Curating Conversation Threads: Expansion, Focus, Volume, Re-entry. In *Proceedings of WSDM*, 2013.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of ICLR*, 2014.

Jiajun Bao, Junjie Wu, Yiming Zhang, Eshwar Chandrasekharan, and David Jurgens. Conversations Gone Alright: Quantifying and Predicting Prosocial Outcomes in Online Conversations. In *Proceedings of WWW*, 2021.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, June 2020.

Cristina Bicchieri. *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. Oxford University Press, December 2016.

Matt Billings and Leon A. Watts. Understanding dispute resolution online: Using text to reflect personal and substantive issues in conflict. In *Proceedings of CHI*, 2010.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec,

Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models. Technical report, Center for Research on Foundation Models, July 2022.

Paolo Bory. Deep new: The shifting narratives of artificial intelligence from Deep Blue to AlphaGo. *Convergence*, 25(4), August 2019.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. Finding Microaggressions in the Wild: A Case for Locating Elusive Phenomena in Social Media Posts. In *Proceedings of EMNLP*, 2019.

Johanna Brewer, Morgan Romine, and T. L. Taylor. Inclusion at Scale: Deploying a Community-Driven Moderation Intervention on Twitch. In *Proceedings of DIS*, 2020.

Arthur C. Brooks. Trolls Aren't Like the Rest of Us. *The Atlantic*, March 2022.

Penelope Brown and Stephen C. Levinson. *Politeness: Some Universals in Language Usage*. Cambridge University Press, 1987.

Moira Burke and Robert Kraut. Mind Your Ps and Qs: The Impact of Politeness and Rudeness in Online Communities. In *Proceedings of CSCW*, 2008.

Anais Cadilhac, Nicholas Asher, Farah Benamara, and Alex Lascarides. Grounding Strategic Conversation: Using Negotiation Dialogues to Predict Trades in a Win-Lose Game. In *Proceedings of EMNLP*, 2013.

Jie Cai and Donghee Yvette Wohn. What are Effective Strategies of Handling Harassment on Twitch? Users' Perspectives. In *Proceedings of CSCW*, 2019.

Jie Cai, Donghee Yvette Wohn, and Mashael Almoqbel. Moderation Visibility: Mapping the Strategies of Volunteer Moderators in Live Streaming Micro Communities. In *Proceedings of IMX*, 2021.

Yang Trista Cao, Lovely-Frances Domingo, Sarah Ann Gilbert, Michelle Mazurek, Katie Shilton, and Hal Daumé III. Toxicity Detection is NOT all you Need: Measuring the Gaps to Supporting Volunteer Content Moderators, November 2023.

Elinor Carmi. Sonic Publics| The Hidden Listeners: Regulating the Line from Telephone Operators to Content Moderators. *International Journal of Communication*, 13(0), January 2019.

Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. #thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. In *Proceedings of CSCW*, 2016.

Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. You Can't Stay Here: The

Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. In *Proceedings of CSCW*, 2017.

Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. In *Proceedings of CSCW*, 2018.

Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. Crossmod: A Cross-Community Learning-based System to Assist Reddit Moderators. In *Proceedings of CSCW*, 2019.

Jonathan P. Chang and Cristian Danescu-Niculescu-Mizil. Trajectories of Blocked Community Members: Redemption, Recidivism and Departure. In *Proceedings of WWW*, 2019a.

Jonathan P. Chang and Cristian Danescu-Niculescu-Mizil. Trouble on the Horizon: Forecasting the Derailment of Online Conversations as they Develop. In *Proceedings of EMNLP*, 2019b.

Jonathan P Chang, Justin Cheng, and Cristian Danescu-Niculescu-Mizil. Don't Let Me Be Misunderstood:Comparing Intentions and Perceptions in Online Discussions. In *Proceedings of WWW*, 2020a.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. ConvoKit: A Toolkit for the Analysis of Conversations. In *Proceedings of SIGDIAL*, 2020b.

Jonathan P. Chang, Charlotte Schluger, and Cristian Danescu-Niculescu-Mizil. Thread With Caution: Proactively Helping Users Assess and Deescalate Tension in Their Online Discussions. In *Proceedings of CSCW*, 2022.

Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. Measuring #GamerGate: A Tale of Hate, Sexism, and Bullying. In *Proceedings of WWW*, 2017.

Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. Antisocial Behavior in Online Discussion Communities. In *Proceedings of ICWSM*, 2015.

Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In *Proceedings of CSCW*, 2017.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of EMNLP*, 2014.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. CONAN - COunter NArratives through Nichesourcing: A Multilingual Dataset of Responses to Fight Online Hate Speech. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of ACL*, 2019.

Herbert Clark and Dale Schunk. Polite responses to polite requests. *Cognition*, 8(2), 1980.

Herbert H. Clark. Responding to indirect speech acts. *Cognitive Psychology*, 11 (4), October 1979.

Benjamin Collier and Julia Bear. Conflict, Criticism, or Confidence: An Empirical Examination of the Gender Gap in Wikipedia Contributions. In *Proceedings of CSCW*, 2012.

Emily I. M. Collins, Anna L. Cox, Jon Bird, and Daniel Harrison. Social networking use and RescueTime: The issue of engagement. In *Proceedings of UbiComp Adjunct*, 2014.

Sunny Consolvo, David W. McDonald, Tammy Toscos, Mike Y. Chen, Jon Froehlich, Beverly Harrison, Predrag Klasnja, Anthony LaMarca, Louis LeGrand, Ryan Libby, Ian Smith, and James A. Landay. Activity sensing in the wild: A field trial of ubifit garden. In *Proceedings of CHI*, 2008.

Lewis A Coser. *The Functions of Social Conflict*. Routledge, 1956.

Kate Crawford and Tarleton Gillespie. What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, 18(3), March 2016.

Katherine Alejandra Cross. In the War Between Harassment and Censorship, No One Wins. *Wired*, September 2023.

Jared R. Curhan and Alex Pentland. Thin Slices of Negotiation: Predicting Outcomes From Conversational Dynamics Within the First 5 Minutes. *Journal of Applied Psychology*, 92, May 2007.

Andrew M. Dai and Quoc V. Le. Semi-supervised Sequence Learning. In *Proceedings of NeurIPS*, 2015.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. Echoes of Power: Language Effects and Power Differences in Social Interaction. In *Proceedings of WWW*, 2012.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. A Computational Approach to Politeness with Application to Social Factors. In *Proceedings of ACL*, 2013a.

Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities. In *Proceedings of WWW*, 2013b.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of ICWSM*, 2017.

Carsten KW De Dreu and Laurie R. Weingart. Task versus relationship conflict, team performance, and team member satisfaction: A meta-analysis. *Journal of applied Psychology*, 88(4), 2003.

Daniel Delmonaco, Samuel Mayworm, Hibby Thach, Josh Guberman, Aurelia Augusta, and Oliver L. Haimson. "What are you doing, TikTok" : How Marginalized Social Media Users Perceive, Theorize, and "Prove" Shadowbanning. In *Proceedings of CSCW*, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*, 2019.

Julian Dibbell. A Rape in Cyberspace. *The Village Voice*, October 2005.

Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning, March 2017.

Bryan Dosono and Bryan Semaan. Moderation Practices as Emotional Labor in Sustaining Online Communities: The Case of AAPI Identity Work on Reddit. In *Proceedings of CHI*, 2019.

Natasha Duarte and Emma Llansó. Mixed Messages? The Limits of Automated Social Media Content Analysis. In *Proceedings of FAccT*, 2018.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. In *Proceedings of EMNLP*, 2021.

Thomas Erickson and Wendy A. Kellogg. Social translucence: An approach to designing systems that support social processes. *ACM Transactions on Computer-Human Interaction*, 7(1), March 2000.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and Removing Disparate Impact. In *Proceedings of KDD*, 2015.

Bruce Fraser. Conversational mitigation. *Journal of Pragmatics*, 4(4), August 1980.

Liye Fu, Lillian Lee, and Cristian Danescu-Niculescu-Mizil. When Confidence and Competence Collide: Effects on Online Decision-Making Discussions. In *Proceedings of WWW*, 2017.

Liye Fu, Susan Fussell, and Cristian Danescu-Niculescu-Mizil. Facilitating the Communication of Politeness through Fine-Grained Paraphrasing. In *Proceedings of EMNLP*, 2020.

Michel Galley, Kathleen McKeown, Julia Hirschberg, and Elizabeth Shriberg. Identifying Agreement and Disagreement in Conversational Speech: Use of Bayesian Networks to Model Pragmatic Dependencies. In *Proceedings of ACL*, 2004.

Björn Gambäck and Utpal Kumar Sikdar. Using Convolutional Neural Networks to Classify Hate-Speech. In *Proceedings of ALW*, 2017.

Jianfeng Gao, Michel Galley, and Lihong Li. Neural Approaches to Conversational AI. In *Proceedings of SIGIR*, 2018.

R. Stuart Geiger. Bot-based collective blocklists in Twitter: The counterpublic moderation of harassment in a networked public space. *Information, Communication & Society*, 19(6), June 2016.

R. Stuart Geiger and David Ribes. The work of sustaining order in wikipedia: The banning of a vandal. In *Proceedings of CSCW*, 2010.

Alex Georgakopoulou, Stefan Iversen, and Carsten Stage. Making Memes Count: Platformed Rallying on Reddit. In Alex Georgakopoulou, Stefan Iversen, and Carsten Stage, editors, *Quantified Storytelling: A Narrative Analysis of Metrics on Social Media*. Springer International Publishing, Cham, 2020.

Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L. Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11), November 2021.

Sarah A. Gilbert. "I run the world's largest historical outreach project and it's on a cesspool of a website." Moderating a Public Scholarship Site on Reddit: A Case Study of r/AskHistorians. In *Proceedings of CSCW*, 2020.

Tarleton Gillespie. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press, New Haven, 2018.

Tarleton Gillespie. Content moderation, AI, and the question of scale:. *Big Data & Society*, July 2020.

Tarleton Gillespie, Patricia Aufderheide, Elinor Carmi, Ysabel Gerrard, Robert Gorwa, Ariadna Matamoros-Fernández, Sarah T. Roberts, Aram Sinnreich, and Sarah Myers West. Expanding the debate about content moderation:

Scholarly research agendas for the coming policy debates. *Internet Policy Review*, 9(4), October 2020.

Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining Explanations: An Overview of Interpretability of Machine Learning. In *Proceedings of the Conference on Data Science and Advanced Analytics*, 2018.

Codruta Girlea, Roxana Girju, and Eyal Amir. Psycholinguistic Features for Deceptive Role Detection in Werewolf. In *Proceedings of NAACL*, 2016.

Erving Goffman. On Face-Work: An Analysis of Ritual Elements in Social Interaction. *Psychiatry*, 18(3), August 1955.

Joseph K. Goodman, Cynthia E. Cryder, and Amar Cheema. Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples. *J. Behav. Decis. Mak.*, 26(3), 2013.

Robert Gorwa. The platform governance triangle: Conceptualising the informal regulation of online content. *Internet Policy Review*, 8(2), June 2019.

Robert Gorwa, Reuben Binns, and Christian Katzenbach. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1), January 2020.

James Grimmelmann. The Virtues of Moderation. *Yale Journal of Law and Technology*, 17(1), September 2015.

David Gurzick, Kevin F. White, Wayne G. Lutters, and Lee Boot. A view from Mount Olympus: The impact of activity tracking tools on the character and practice of moderation. In *Proceedings of GROUP*, 2009.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. Before Name-Calling: Dynamics and Triggers of Ad Hominem Fallacies in Web Argumentation. In *Proceedings of NAACL*, 2018.

Aaron Halfaker, Aniket Kittur, and John Riedl. Don't bite the newbies: How reverts affect the quantity and quality of Wikipedia work. In *Proceedings of WikiSym*, 2011a.

Aaron Halfaker, Bryan Song, D. Alex Stuart, Aniket Kittur, and John Riedl. NICE: Social Translucence Through UI Intervention. In *Proceedings of WikiSym*, 2011b.

Aaron Halfaker, R. Stuart Geiger, Jonathan T. Morgan, and John Riedl. The Rise and Decline of an Open Collaboration System. *American Behavioral Scientist*, 57(5), May 2013.

Xiaochuang Han and Yulia Tsvetkov. Fortifying Toxic Speech Detectors Against Veiled Toxicity. In *Proceedings of EMNLP*, 2020.

Amy A. Hasinoff and Nathan Schneider. From Scalability to Subsidiarity in Addressing Online Harm. *Social Media + Society*, 8(3), July 2022.

Bing He, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar. Racism is a virus: Anti-asian hate and counterspeech in social media during the COVID-19 crisis. In *Proceedings of ASONAM*, 2022.

Christophe Henner and Maria Sefidari. Wikimedia Foundation Board on healthy Wikimedia community culture, inclusivity, and safe spaces – Wikimedia Blog. https://blog.wikimedia.org/2016/12/08/board-culture-inclusivity-safe-spaces/, 2016.

Jack Hessel and Lillian Lee. Something's Brewing! Early Prediction of Controversy-causing Posts from Discussion Features. In *Proceedings of NAACL*, 2019.

Francis Heylighen and Jean-Marc Dewaele. Formality of Language: Definition, measurement and behavioral determinants. Technical report, Center "Leo Apostel", Free University of Brussels, 1999.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. Analyzing the Semantic Types of Claims and Premises in an Online Persuasive Forum. In *Proceedings of the 4th Workshop on Argument Mining*, 2017.

Eric Holgate, Isabel Cachola, Daniel Preotiuc-Pietro, and Junyi Jessy Li. Why Swear? Analyzing and Inferring the Intentions of Vulgar Expressions. In *Proceeding of EMNLP*, 2018.

Jeremy Howard and Sebastian Ruder. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of ACL*, 2018.

Yilun Hua, Nicholas Chernogor, Yuzhe Gu, Seoyeon Julie Jeong, Miranda Luo, and Crisitan Danescu-Niculescu-Mizil. How Did We Get Here? Summarizing Conversation Dynamics. In *Proceedings of NAACL*, 2024.

Yiqing Hua, Cristian Danescu-Niculescu-Mizil, Dario Taraborelli, Nithum Thain, Jeffery Sorensen, and Lucas Dixon. WikiConv: A Corpus of the Complete Conversational History of a Large Online Collaborative Community. In *Proceedings of EMNLP*, 2018.

Karen Huang, Michael Yeomans, Alison Wood Brooks, Julia Minson, and

Francesca Gino. It doesn't hurt to ask: Question-asking increases liking. *J. Pers. Soc. Psychol.*, 113(3), September 2017.

Axel Hübler. *Understatements and Hedges in English*. John Benjamins Publishing Company, Amsterdam Philadelphia, January 1983.

Krithika Jagannath, Katie Salen, and Petr Slovàk. "(We) Can Talk It Out...": Designing for Promoting Conflict-Resolution Skills in Youth on a Moderated Minecraft Server. In *Proceedings of CSCW*, 2020.

Piotr Janiszewski, Mateusz Lango, and Jerzy Stefanowski. Time Aspect in Making an Actionable Prediction of a Conversation Breakdown. In *Proceedings of KDD*, 2021.

Shagun Jhaver, Pranil Vora, and Amy Bruckman. Designing for Civil Conversations: Lessons Learned from ChangeMyView. Technical Report, Georgia Institute of Technology, December 2017.

Shagun Jhaver, Larry Chan, and Amy Bruckman. The view from the other side: The border between controversial speech and harassment on Kotaku in Action. *First Monday*, 23(2), February 2018a.

Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. Online Harassment and Content Moderation: The Case of Blocklists. *ACM Trans. Comput-Hum. Interact.*, 25(2), March 2018b.

Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. "Did You Suspect the Post Would be Removed?": Understanding User Reactions to Content Removals on Reddit. In *Proceedings of CSCW*, 2019a.

Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. Human-Machine

Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Trans. Comput-Hum. Interact.*, 26(5), July 2019b.

Shagun Jhaver, Christian Boylston, Diyi Yang, and Amy Bruckman. Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter. In *Proceedings of CSCW*, 2021.

Yohan Jo, Shivani Poddar, Byungsoo Jeon, Qinlan Shen, Carolyn P. Rosé, and Graham Neubig. Attentive Interaction Model: Modeling Changes in View in Argumentation. In *Proceedings of NAACL*, 2018.

Amy Johnson. The Multiple Harms of Sea Lions. In *Perspectives on Harmful Speech Online.* Berkman Klein Center for Internet & Society., 2017.

Brihi Joshi, Ziyi Liu, Sahana Ramnath, Aaron Chan, Zhewei Tong, Shaoliang Nie, Qifan Wang, Yejin Choi, and Xiang Ren. Are Machine Rationales (Not) Useful to Humans? Measuring and Improving Human Utility of Free-Text Rationales. In *Proceedings of ACL*, 2023.

Jukka Jouhki, Epp Lauk, Maija Penttinen, Niina Sormanen, and Turo Uskali. Facebook's Emotional Contagion Experiment as a Challenge to Research Ethics. *Media and Communication*, 4(4), October 2016.

David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. A Just and Comprehensive Strategy for Using NLP to Address Online Abuse. In *Proceedings of ACL*, 2019.

Matthew Katsaros, Kathy Yang, and Lauren Fratamico. Reconsidering Tweets: Intervening during Tweet Creation Decreases Offensive Content. In *Proceedings of ICWSM*, 2022.

Christian Katzenbach. "AI will fix this" – The Technical, Discursive, and Political Turn to AI in Governing Communication. *Big Data & Society*, 8(2), July 2021.

Joseph M. Kayany. Contexts of uninhibited online behavior: Flaming in social newsgroups on usenet. *Journal of the American Society for Information Science*, 49(12), January 1998.

Yova Kementchedjhieva and Anders Sogaard. Dynamic Forecasting of Conversation Derailment. In *Proceedings of EMNLP*, 2021.

Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. Regulating behavior in online communities. In Paul Resnick and Robert Kraut, editors, *Building Successful Online Communities: Evidence-Based Social Design*. MIT Press, 2012.

Aniket Kittur and Robert E. Kraut. Harnessing the Wisdom of Crowds in Wikipedia: Quality Through Coordination. In *Proceedings of CSCW*, 2008.

Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. He says, she says: Conflict and coordination in Wikipedia. In *Proceedings of CHI*, 2007.

Aniket Kittur, Ed H. Chi, and Bongwon Suh. What's in Wikipedia? mapping topics and conflict using socially annotated category structure. In *Proceedings of CHI*, 2009.

Olivier Klein, Stéphane Doyen, Christophe Leys, Pedro A. Magalhães de Saldanha da Gama, Sarah Miller, Laurence Questienne, and Axel Cleeremans. Low Hopes, High Expectations: Expectancy Effects and the Replicability of Behavioral Experiments. *Perspectives on Psychological Science*, 7(6), November 2012.

Anne Kohlbrenner, Ben Kaiser, Kartikeya Kandula, Rebecca Weiss, Jonathan Mayer, Ted Han, and Robert Helmer. Rally and WebScience: A Platform and Toolkit for Browser-Based Research on Technology and Society Problems, November 2022.

Yubo Kou. Toxic Behaviors in Team-Based Competitive Gaming: The Case of League of Legends. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, 2020.

Yubo Kou and Xinning Gui. Flag and Flaggability in Automated Moderation: The Case of Reporting Toxic Behavior in an Online Game Community. In *Proceedings of CHI*, 2021.

Geza Kovacs, Zhengxuan Wu, and Michael S. Bernstein. Rotating Online Behavior Change Interventions Increases Effectiveness But Also Increases Attrition. In *Proceedings of CSCW*, 2018.

Paul Krebs, James O. Prochaska, and Joseph S. Rossi. A meta-analysis of computer-tailored interventions for health behavior change. *Preventive Medicine*, 51(3), September 2010.

Travis Kriplean, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett. Supporting reflective public thought with considerit. In *Proceedings of CSCW*, 2012a.

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 CHI*, 2012b.

Vinodh Krishnan and Jacob Eisenstein. "You're Mr. Lebowski, I'm the Dude":

Inducing Address Term Formality in Signed Social Networks. In Rada Mihalcea, Joyce Chai, and Anoop Sarkar, editors, *Proceedings of NAACL*, 2015.

Ravi Kumar, Mohammad Mahdian, and Mary McGlohon. Dynamics of conversations. In *Proceedings of KDD*, 2010.

Srijan Kumar, Justin Cheng, Jure Leskovec, and V.S. Subrahmanian. An Army of Me: Sockpuppets in Online Discussion Communities. In *Proceedings of WWW*, 2017.

Srijan Kumar, William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Community Interaction and Conflict on the Web. In *Proceedings of WWW*, 2018.

Haewoon Kwak, Jeremy Blackburn, and Seungyeop Han. Exploring Cyberbullying and Other Toxic Behavior in Team Competition Online Games. In *Proceedings of CHI*, 2015.

Robin T. Lakoff. *The Logic of Politeness: Minding Your P's and Q's*. Chicago Linguistic Society, 1973.

Cliff Lampe and Paul Resnick. Slash(dot) and burn: Distributed moderation in a large online conversation space. In *Proceedings of CHI*, 2004.

Sarah Ita Levitan, Angel Maredia, and Julia Hirschberg. Linguistic Cues to Deception and Perceived Deception in Interview Dialogues. In *Proceedings of NAACL*, 2018.

Renee Li, Pavitthra Pandurangan, Hana Frluckaj, and Laura Dabbish. Code of Conduct Conversations in Open Source Software Projects on Github. In *Proceedings of CSCW*, 2021.

Zhenhao Li, Marek Rei, and Lucia Specia. Multimodal Conversation Modelling for Topic Derailment Detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022.

Bing Liu, Minqing Hu, and Junsheng Cheng. Opinion observer: Analyzing and comparing opinions on the Web. In *Proceedings of WWW*, 2005.

Ping Liu, Joshua Guberman, Libby Hemphill, and Aron Culotta. Forecasting the Presence and Intensity of Hostility on Instagram Using Linguistic and Social Features. In *Proceedings of ICWSM*, 2018.

Claudia (Claudia Wai Yu) Lo. *When All You Have Is a Banhammer : The Social and Communicative Work of Volunteer Moderators*. Thesis, Massachusetts Institute of Technology, 2018.

Josh Magnus Ludan, Qing Lyu, Yue Yang, Liam Dugan, Mark Yatskar, and Chris Callison-Burch. Interpretable-by-Design Text Classification with Iteratively Generated Concept Bottleneck, October 2023.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of EMNLP*, 2015.

Lena Mamykina, Bella Manoim, Manas Mittal, George Hripcsak, and Björn Hartmann. Design lessons from the fastest Q&A site in the west. In *Proceedings of CHI*, 2011.

Andrew Marantz. Reddit and the Struggle to Detoxify the Internet. *The New Yorker*, March 2018.

Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. Thou

Shalt Not Hate: Countering Online Hate Speech. *Proceedings of the International AAAI Conference on Web and Social Media*, 13, July 2019.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. Learned in Translation: Contextualized Word Vectors. In *Proceedings of NeurIPS*, 2017.

Heidi McKee. "YOUR VIEWS SHOWED TRUE IGNORANCE!!!": (Mis)Communication in an online interracial discussion forum. *Computers and Composition*, 19(4), December 2002.

Julia Mendelsohn, Ronan Le Bras, Yejin Choi, and Maarten Sap. From Dogwhistles to Bullhorns: Unveiling Coded Rhetoric with Language Models. In *Proceedings of ACL*, 2023.

Ryan M. Milner. *The World Made Meme: Discourse and Identity in Participatory Media*. PhD thesis, University of Kansas, August 2012.

Jonathan T. Morgan and Aaron Halfaker. Evaluating the impact of the Wikipedia Teahouse on newcomer socialization and retention. In *Proceedings of OpenSym*, 2018.

Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen Carley. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. *Proceedings of the International AAAI Conference on Web and Social Media*, 7(1), 2013.

Aske Mottelson and Kasper Hornbæk. Virtual reality studies outside the laboratory. In *Proceedings of VRST*, 2017.

Annalee Newitz. Opinion | We Forgot About the Most Important Job on the Internet. *The New York Times*, March 2020.

Austin Lee Nichols and Jon K. Maner. The Good-Subject Effect: Investigating Participant Demand Characteristics. *The Journal of General Psychology*, 135(2), April 2008.

Vlad Niculae and Cristian Danescu-Niculescu-Mizil. Conversational Markers of Constructive Discussions. In *Proceedings of NAACL*, 2016.

Vlad Niculae, Srijan Kumar, Jordan Boyd-Graber, and Cristian Danescu-Niculescu-Mizil. Linguistic Harbingers of Betrayal: A Case Study on an Online Strategy Game. In *Proceedings of ACL*, 2015.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive Language Detection in Online User Content. In *Proceedings of WWW*, 2016.

Daniel J. O'Keefe. Standpoint Explicitness and Persuasive Effect: A Meta-Analytic Review of the Effects of Varying Conclusion Articulation in Persuasive Messages. *Argumentation and Advocacy*, 34(1), June 1997.

Daniel J. O'Keefe. Justification Explicitness and Persuasive Effect: A Meta-Analytic Review of the Effects of Varying Support Articulation in Persuasive Messages. *Argumentation and Advocacy*, 35(2), September 1998.

Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush Varshney. The Effect of Extremist Violence on Hateful Speech Online. In *Proceedings of ICWSM*, 2018.

Marco Ortu, Bram Adams, Giuseppe Destefanis, Parastou Tourani, Michele Marchesi, and Roberto Tonelli. Are Bullies More Productive? Empirical Study of Affectiveness vs. Issue Fixing Time. In *2015 IEEE/ACM 12th Working Conference on Mining Software Repositories*, 2015.

Orestis Papakyriakopoulos, Severin Engelmann, and Amy Winecoff. Upvotes? Downvotes? No Votes? Understanding the relationship between reaction mechanisms and political discourse on Reddit. In *Proceedings of CHI*, 2023.

Deokgun Park, Simranjit Sachar, Nicholas Diakopoulos, and Niklas Elmqvist. Supporting Comment Moderators in Identifying High Quality Online News Comments. In *Proceedings of CHI*, 2016.

Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing Gender Bias in Abusive Language Detection. In *Proceedings of EMNLP*, 2018.

Nancy Paterson. Walled gardens: The new shape of the public internet. In *Proceedings of iConference*, 2012.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. Deep Learning for User Comment Moderation. In *Proceedings of ALO*, 2017a.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. Deeper Attention to Abusive User Content Moderation. In *Proceedings of EMNLP*, 2017b.

James W. Pennebaker, Cindy K. Chung, Joey Frazee, Gary M. Lavergne, and David I. Beaver. When Small Words Foretell Academic Success: The Case of College Admissions Essays. *PLOS ONE*, 2014.

Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, Yao Xiao, C. J. Linton, and Mihai Burzo. Verbal and Nonverbal Clues for Real-life Deception Detection. In *Proceedings of EMNLP*, 2016.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep Contextualized Word Representations. In *Proceedings of NAACL*, 2018.

Peter Potash and Anna Rumshisky. Towards Debate Automation: A Recurrent Model for Predicting Debate Winners. In *Proceedings of EMNLP*, 2017.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-training. Technical report, OpenAI, 2018.

Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. The Fallacy of AI Functionality. In *Proceedings of FAccT*, 2022.

Katharina Reinecke and Krzysztof Z. Gajos. LabintheWild: Conducting Large-Scale Online Experiments With Uncompensated Samples. In *Proceedings of CSCW*, 2015.

Manoel Horta Ribeiro, Pedro H. Calais, Yuri A. Santos, Virgílio A. F. Almeida, and Wagner Meira Jr. Characterizing and Detecting Hateful Users on Twitter. In *Proceedings of ICWSM*, 2018.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of ACL*, 2020.

Annika Richterich. 'Karma, Precious Karma!' Karmawhoring on Reddit and the Front Page's Econometrisation. *Journal of Peer Production*, 4(1), 2014.

Alan Ritter, Colin Cherry, and Bill Dolan. Unsupervised Modeling of Twitter Conversations. In *Proceedings of NAACL*, 2010.

Sarah T Roberts. *Behind the Screen: The Hidden Digital Labor of Commercial Content Moderation*. PhD thesis, University of Illinois at Urbana-Champagne, 2014.

Paul R. Rosenbaum. *Design of Observational Studies*. Springer, New York, 2010.

Sara Rosenthal and Kathleen McKeown. I Couldn't Agree More: The Role of Conversational Structure in Agreement and Disagreement Detection in Online Discussions. In *Proceedings of SIGDIAL*, 2015.

Donald B. Rubin. The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Stat. Med.*, 26(1), January 2007.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of ACL*, 2019.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *Proceedings of ACL*, 2020.

Martin Saveski, Brandon Roy, and Deb Roy. The Structure of Toxic Conversations on Twitter. In *Proceedings of WWW*, 2021.

Charlotte Schluger, Jonathan P. Chang, Cristian Danescu-Niculescu-Mizil, and Karen Levy. Proactive Moderation of Online Discussions: Existing Practices and the Potential for Algorithmic Support. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2), November 2022.

Joseph Seering. Reconsidering Self-Moderation: The Role of Research in Supporting Community-Based Models for Online Content Moderation. In *Proceedings of CSCW*, 2020.

Joseph Seering and Sanjay R. Kairam. Who Moderates on Twitch and What Do They Do? Quantifying Practices in Community Moderation on Twitch. *Proceedings of the ACM on Human-Computer Interaction*, 7(GROUP), December 2022.

Joseph Seering, Robert Kraut, and Laura Dabbish. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of CSCW*, 2017.

Joseph Seering, Tianmi Fang, Luca Damasco, Mianhong 'Cherie' Chen, Likang Sun, and Geoff Kaufman. Designing User Interface Elements to Improve the Quality and Civility of Discourse in Online Commenting Behaviors. In *Proceedings of CHI*, 2019a.

Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. Moderator engagement and community development in the age of algorithms. *New Media & Society*, 21(7), July 2019b.

Joseph Seering, Geoff Kaufman, and Stevie Chancellor. Metaphors in moderation. *New Media & Society*, October 2020.

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *Proceedings of AAAI*, 2016.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. In *Proceedings of AAAI*, 2017.

Farhana Shahid, Dhruv Agarwal, and Aditya Vashistha. 'One Style Does Not Regulate All': Moderation Practices in Public and Private WhatsApp Groups, January 2024.

Cuihua Shen, Qiusi Sun, Taeyoung Kim, Grace Wolff, Rabindra Ratan, and

Dmitri Williams. Viral vitriol: Predictors and contagion of online toxicity in World of Tanks. *Computers in Human Behavior*, 108, July 2020.

Amit Sheth, Valerie L. Shalin, and Ugur Kursuncu. Defining and detecting toxicity on social media: Context and knowledge are key. *Neurocomputing*, 490, June 2022.

Anthony Sicilia, Hyunwoo Kim, Khyathi Raghavi Chandu, Malihe Alikhani, and Jack Hessel. Deal, or no deal (or who knows)? Forecasting Uncertainty in Conversations using Large Language Models, February 2024.

Vivek K. Singh, Marie L. Radford, Qianjia Huang, and Susan Furrer. "They basically like destroyed the school one day": On Newer App Features and Cyberbullying in Schools. In *Proceedings of CSCW*, 2017.

Ana Smith. *Leveraging Context Documents for Social Natural Language Processing*. PhD thesis, Cornell University, United States – New York, 2023.

Marina Sokolova, Vivi Nastase, and Stan Szpakowicz. The Telling Tail: Signals of Success in Electronic Negotiation Texts. In *Proceedings of IJCNLP*, 2008.

Sara Owsley Sood, Elizabeth F. Churchill, and Judd Antin. Automatic identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology*, 63(2), February 2012.

Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. A Hierarchical Recurrent Encoder-Decoder for Generative Context-Aware Query Suggestion. In *Proceedings of CIKM*, 2015a.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. A Neural

Network Approach to Context-Sensitive Generation of Conversational Responses. In *Proceedings of NAACL*, 2015b.

Kumar Srinivasan, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan. Content Removal as a Moderation Strategy: Compliance and Other Outcomes in the ChangeMyView Community. In *Proceedings of CSCW*, 2019.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to Sequence Learning with Neural Networks. In *Proceedings of NeurIPS*, 2014.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu, and Lillian Lee. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions. In *Proceedings of WWW*, 2016.

Yla R Tausczik and James W Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *J. Lang. Soc. Psychol.*, 29(1), 2010.

Samuel Hardman Taylor, Dominic DiFranzo, Yoon Hyung Choi, Shruti Sannon, and Natalya N. Bazarova. Accountability and Empathy by Design: Encouraging Bystander Intervention to Cyberbullying on Social Media. In *Proceedings of CSCW*, 2019.

Nithum Thain, Lucas Dixon, and Ellery Wulczyn. Wikipedia Talk Labels: Toxicity, February 2017.

Kaitlyn Tiffany. Inside R/Relationships, the Unbearably Human Corner of Reddit. *The Atlantic*, October 2019.

Kal Turnbull. "That's Bullshit" – Rude Enough for Removal? A Multi-Mod Perspective, March 2018.

U.S. House of Representatives. *Joint Hearing on 'Fostering a Healthier Internet to Protect Consumers'*. House Committee on Energy and Commerce, October 2019. URL `https://catalog.gpo.gov/F/?func=direct&doc_number=001167713&format=999`.

Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. Identifying Women's Experiences With and Strategies for Mitigating Negative Effects of Online Harassment. In *Proceedings of CSCW*, 2017.

Svitlana Volkova and Eric Bell. Identifying Effective Signals to Predict Deleted and Suspended Accounts on Twitter across Languages. In *Proceedings of ICWSM*, 2017.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. Retrieval of the Best Counterargument without Prior Topic Knowledge. In *Proceedings of ACL*, 2018.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2), 2018.

Lu Wang and Claire Cardie. A Piece of My Mind: A Sentiment Analysis Approach for Online Dispute Detection. In *Proceedings of ACL*, 2014.

Lu Wang, Nick Beauchamp, Sarah Shugars, and Kechen Qin. Winning on the Merits: The Joint Effects of Content and Style on Debate Outcomes. *Transactions of the Association for Computational Linguistics*, 5, 2017.

William Warner and Julia Hirschberg. Detecting Hate Speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media*, Montreal, Canada, 2012.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent Abilities of Large Language Models. In *Proceedings of TMLR*, 2022.

Zhongyu Wei, Yang Liu, and Yi Li. Is This Post Persuasive? Ranking Argumentative Comments in the Online Forum. In *Proceedings of ACL*, 2016.

Galen Weld, Amy X. Zhang, and Tim Althoff. What Makes Online Communities 'Better'? Measuring Values, Consensus, and Conflict across Thousands of Subreddits. In *Proceedings of ICWSM*, 2022.

Galen Weld, Leon Leibmann, Amy X. Zhang, and Tim Althoff. Perceptions of Moderators as a Large-Scale Measure of Online Community Governance, January 2024.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. Detection of Abusive Language: The Problem of Biased Datasets. In *Proceedings of NAACL*, 2019.

Wikimedia Support and Safety Team. Harassment Survey, 2015. URL `https://upload.wikimedia.org/wikipedia/commons/5/52/Harassment_Survey_2015_-_Results_Report.pdf`.

Sam Witteveen and Martin Andrews. Paraphrasing with Large Language Models. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, 2019.

Donghee Yvette Wohn. Volunteer Moderators in Twitch Micro Communities:

How They Get Involved, the Roles They Play, and the Emotional Labor They Experience. In *Proceedings of CHI*, 2019.

Austin P. Wright, Omar Shaikh, Haekyu Park, Will Epperson, Muhammed Ahmed, Stephane Pinel, Diyi Yang, and Duen Horng Chau. RECAST: Interactive Auditing of Automatic Toxicity Detection Models. In *Proceedings of CSCW*, 2021.

Lucas Wright. Automated Platform Governance Through Visibility and Scale: On the Transformational Power of AutoModerator. *Social Media + Society*, 8 (1), January 2022.

Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development. *IEEE/CAA Journal of Automatica Sinica*, 10(5), May 2023.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of WWW*, 2017.

Diyi Yang, Jiaao Chen, Zichao Yang, Dan Jurafsky, and Eduard Hovy. Let's Make Your Request More Persuasive: Modeling Persuasive Strategies via Semi-Supervised Neural Nets on Crowdfunding Platforms. In *Proceedings of NAACL*, 2019.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a Bottle: Language Model Guided Concept Bottlenecks for Interpretable Image Classification. In *Proceedings of CVPR*, 2023.

Dawei Yin, Zhenzhen Xue, and Liangjie Hong. Detection of Harassment on Web 2.0. In *Proceedings of CAW2.0*, 2009.

Amy Zhang, Bryan Culbertson, and Praveen Paritosh. Characterizing Online Discussion Using Coarse Discourse Sequences. In *Proceedings of ICWSM*, 2017a.

Amy X. Zhang, Grant Hugh, and Michael S. Bernstein. PolicyKit: Building Governance in Online Communities. In *Proceedings of UIST*, 2020a.

Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. Demographics Should Not Be the Reason of Toxicity: Mitigating Discrimination in Text Classifications with Instance Weighting. In *Proceedings of ACL*, 2020b.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. Conversational Flow in Oxford-style Debates. In *Proceedings of NAACL*, 2016.

Justine Zhang, Arthur Spirling, and Cristian Danescu-Niculescu-Mizil. Asking too Much? The Rhetorical Role of Questions in Political Discourse. In *Proceedings of EMNLP*, 2017b.

Justine Zhang, Jonathan P. Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Nithum Thain, Yiqing Hua, and Dario Taraborelli. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of ACL*, 2018a.

Justine Zhang, Cristian Danescu-Niculescu-Mizil, Christina Sauper, and Sean J. Taylor. Characterizing Online Public Discussions Through Patterns of Participant Interactions. In *Proceedings of CSCW*, 2018b.

Jichen Zhu, Antonios Liapis, Sebastian Risi, Rafael Bidarra, and G. Michael Youngblood. Explainable AI for Designers: A Human-Centered Perspective on Mixed-Initiative Co-Creation. In *Proceedings of CIG*, 2018.