Thread With Caution: Proactively Helping Users Assess and Deescalate Tension in Their Online Discussions

JONATHAN P. CHANG, Cornell University, USA
CHARLOTTE SCHLUGER, Cornell University, USA
CRISTIAN DANESCU-NICULESCU-MIZIL, Cornell University, USA

Incivility remains a major challenge for online discussion platforms, to such an extent that even conversations between well-intentioned users can often derail into uncivil behavior. Traditionally, platforms have relied on moderators to—with or without algorithmic assistance—take corrective actions such as removing comments or banning users. In this work we propose a complementary paradigm that directly empowers users by proactively enhancing their awareness about existing tension in the conversation they are engaging in and actively guides them as they are drafting their replies to avoid further escalation.

As a proof of concept for this paradigm, we design an algorithmic tool that provides such proactive information directly to users, and conduct a user study in a popular discussion platform. Through a mixed methods approach combining surveys with a randomized controlled experiment, we uncover qualitative and quantitative insights regarding how the participants utilize and react to this information. Most participants report finding this proactive paradigm valuable, noting that it helps them to identify tension that they may have otherwise missed and prompts them to further reflect on their own replies and to revise them. These effects are corroborated by a comparison of how the participants draft their reply when our tool warns them that their conversation is at risk of derailing into uncivil behavior versus in a control condition where the tool is disabled. These preliminary findings highlight the potential of this user-centered paradigm and point to concrete directions for future implementations.

CCS Concepts: • **Human-centered computing** → *Interactive systems and tools*; **Collaborative and social computing systems and tools**; • **Computing methodologies** → **Natural language processing**.

Additional Key Words and Phrases: Incivility, online discussions, proactive intervention, tension, forecasting, antisocial behavior, prosocial intervention.

ACM Reference Format:
Jonathan P. Chang, Charlotte Schluger, and Cristian Danescu-Niculescu-Mizil. 2022. Thread With Caution: Proactively Helping Users Assess and Deescalate Tension in Their Online Discussions. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 545 (November 2022), 37 pages. https://doi.org/10.1145/3555603

1 INTRODUCTION

Incivility remains an important issue in online discussion platforms [11], hindering the exchange of ideas [1] and taking a significant emotional toll on the participants [2, 43]. Traditionally, platforms attempt to address this problem through reactive moderation, in which volunteers from within the community [67] or professionals employed by the platform operator [28] aim to identify and remove "bad actors" and "objectionable content". Substantial efforts are focusing on scaling up this

Authors' addresses: Jonathan P. Chang, jpc362@cornell.edu, Cornell University, Ithaca, NY, USA, 14850; Charlotte Schluger, jes543@cornell.edu, Cornell University, Ithaca, NY, USA, 14850; Cristian Danescu-Niculescu-Mizil, cristian@cs.cornell.edu, Cornell University, Ithaca, NY, USA, 14850.

2573-0142/2022/11-ART545 $15.00
https://doi.org/10.1145/3555603

paradigm through automation or algorithmic assistance, an enterprise which has proven to be both technically and ethically challenging [29, 32, 33, 46].

This common paradigm, however, does not account for the fact that uncivil behavior in discussion platforms is not solely the product of "bad actors"—who are generally a minority within their communities [54]—but can instead often emerge from ordinary users when they find themselves in particularly heated or tense situations [15]. 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 [27, 72, 77]. Starting from this viewpoint, in this work we propose a complementary paradigm that directly empowers such well-intentioned users to proactively avoid escalating tense situations. It does so by informing them when the conversation they are engaging in is at risk of derailing into incivility, i.e., enhancing their *risk awareness*.

To test whether this proactive risk awareness paradigm is feasible in a real world setting—and thus avoid artifacts of laboratory or crowdsourced studies [31, 64, 74]—we conduct a user study in a popular online discussion platform, ChangeMyView. Executing this real-world user study 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 both automatically assess the risk of derailment for ongoing discussions in real time and inform the participants about this risk 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 a backend algorithmic scoring system powered by a recent natural language processing methodology—conversational forecasting [14, 57, 81]—and a frontend browser plugin that serves the resulting information to users directly on the ChangeMyView webpage (see the Video Figure).[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 report that 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.

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

[1] A link to the Video Figure can be found at https://www.cs.cornell.edu/~cristian/Thread_With_Caution.html.

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 paper 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.

2 BACKGROUND AND RELATED WORK

Every online discussion platform must eventually reckon with the problem of incivility and other undesirable behaviors, leading to Gillespie's assertion that "all platforms moderate" [28]. Yet the *specific* practices for tackling this problem vary greatly across different platforms, and in turn a rich body of prior literature on moderation has sought to categorize and compare these different strategies. To better contextualize our work within the landscape of moderation, we first synthesize prior work's categorization of moderation practices, identifying two key axes of design: *who* does the work of moderation and *when* this work happens. Then, we explain where our proposed paradigm falls within this typology.

2.1 Actors Involved in Moderation: Who Does the Work?

Today, many online platforms take a *platform-driven* approach to moderation, where platform operators directly employ or contract workers to review potentially objectionable content and remove it if needed [28]. 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 [11]. 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 [29], and results in a high workload and stress for the moderators [66].

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 [21, 56]. As of late, community-driven moderation has seen a renewed surge in interest in light of the shortcomings of platform-driven moderation [6, 67], and it remains the method of choice in smaller, interest-specific communities—for example, Twitch livestream communities [9, 58] and the topical groups on Reddit known as "subreddits" [12, 22, 27]. Community-driven moderation practices can be further subdivided as roughly falling into two categories: volunteer moderators and end-user tools.

One common approach to community-driven moderation 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 [22, 25, 58, 67, 78]. 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 [67], and their inside knowledge of community norms and dynamics can help them negotiate harder, more nuanced disputes [12, 75]. 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 [22, 78].

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 community-driven moderation strategies: *end-user tools*. In contrast to the previously described centralized strategies, end-user tools are accessible to *all* community members, distributing the work of content moderation across the entire community [47, 67]. While the risk of misuse necessarily implies that end-user tools cannot be as authoritative as moderators' tools (e.g., end users should not have the ability to remove someone else's content), platforms have managed to innovate various softer approaches that have met with some success. A particularly common end-user tool is the ability 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 [12, 56, 59]. An even softer end-user tool is the personalized *blocklist* [24, 44], which shifts the goal from removing objectionable content from the platform to simply removing it from an individual user's feed.

2.2 When Does Moderation Work Happen?

While the moderation strategies we have described so far are quite varied, they all have one thing in common: they are designed to deal with uncivil content that has *already* been created. While dealing with uncivil content after-the-fact is better than doing nothing, some have argued that a more effective way to protect online communities from harm is to reduce the amount of uncivil content that gets created in the first place [33, 47]. Strategies designed to achieve this have been referred to as *proactive* moderation [10, 58, 67, 70], as a contrast to the previously described after-the-fact strategies that are referred to as *reactive* moderation.

In current practice, proactive moderation is largely the realm of volunteer moderators, whose combination of authority and connection to the community put them in a unique position to engage in social strategies to guide user behavior towards healthier interactions [69, 71]. Such strategies include educating users about the community's rules [9, 71], publicly modeling good behavior [40, 70], or mediating disputes before they can get out of hand [4]. In interviews, volunteer moderators have indicated that they see this as just another part of their job, describing the work using metaphors like "teacher" and "facilitator" [69].

Where our paradigm fits: User-facing proactive interventions. The proactive strategies we have discussed thus far involve moderator actions. But as we have previously seen, when it comes to *reactive* strategies, there is room for end users to play a role alongside moderators in the broader landscape of moderation. Does the same hold true for proactive strategies?

This question has driven a recently emerging line of research that looks at how platform design could directly steer end users towards more prosocial behaviors, in the form of user-facing *interventions*. Various intervention strategies have been attempted: some work explicitly encourages users to take particular actions, such as reflecting more deeply on comments they have read [51, 52]; others operate on a more subconscious level, by asking users to complete tasks prior to commenting

that might prime them to be more prosocial [68, 74]; and still others simply aim to provide users with additional information, such as the fact that the other user they are engaging with is new to the community [35], in hopes that this additional information might influence their subsequent behavior.

While experiments with these interventions have shown success, the authors of one such system, Taylor et al. [74], 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 [16, 49, 50]. 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."

Our present work adds to the ongoing research on proactive intervention design by exploring Taylor et al.'s proposal of targeted, just-in-time interventions. We are inspired by a recent algorithmic development, the emergence of *conversational forecasting* algorithms (Section 3.1.1) that could provide the technical backbone for Taylor et al.'s proposal, automatically identifying interactions in need of intervention by detecting *rising tension* that could lead to incivility in the future. We use this technology to build our own implementation of a proactive intervention system (Section 3.1), and heeding Taylor et al.'s caution about the limitations of lab studies, we describe and execute a plan for evaluating the intervention in a real-world setting (Section 3.2).

3 METHODS

To evaluate the feasibility of our proposed risk awareness paradigm we develop a prototype tool that implements it, ConvoWizard (Section 3.1), and gather both qualitative feedback and quantitative usage data through an IRB-approved user study (Section 3.2). Heeding Taylor et al.'s warnings of the potential inadequacies of laboratory studies in understanding the impact of prosocial interventions, we design our user study to capture how regular users might be affected by the intervention in their everyday online interactions. Prior work throughout the HCI and CSCW space has argued that achieving this goal requires going outside the laboratory and testing the intervention in a real world setting, or "in the wild" [7, 17, 60, 64]. Following this line of work, we set up our user study to involve real users in an actual social media community, namely the Reddit debate forum ChangeMyView. 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 3.1, we explain the technical approach we take to tackling this problem, developing a browser extension that uses established techniques to read the content of conversations taking place on Reddit, algorithmically score the risk level of that content in real time, and extend 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 3.2.2, 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 [20, 23], 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 3.2.1, we explain this approach in more detail.

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 a machine learning model in real time to predict the trajectory of ongoing conversations, relays this information to the browser extension, and logs data for subsequent analysis.

3.1.1 Underlying machine learning model: The conversational forecasting algorithm. To automatically estimate the risk of future incivility in conversations, ConvoWizard leverages a recent Natural Language Processing paradigm, *conversational forecasting*, which trains models to predict future conversational outcomes based on the current state of the conversation [14, 57, 81]. Specifically, ConvoWizard uses CRAFT, a conversational forecasting model that was trained to forecast the future occurrence of uncivil behavior in a conversation [14]. An appealing aspect of CRAFT for our study is that it has previously been trained and evaluated on ChangeMyView, using "natural" labels that came from ChangeMyView moderator actions, specifically removals of comments that violated ChangeMyView's rules against uncivil behavior.[2]

Formally, given a conversation $C = \{c_1, c_2, \ldots, c_n\}$ represented as a series of comments in reply-to order, CRAFT predicts the likelihood that the next comment c_{n+1} will contain uncivil behavior. In other words, $\text{CRAFT}(C) = p(\text{isUncivil}(c_{n+1}))$. In this way, given the current state of a conversation, CRAFT can predict the risk that the next comment that gets posted will exhibit uncivil behavior. Moreover, CRAFT is an online model: when a new comment c_{n+1} is eventually added to the discussion, CRAFT can compute an updated prediction for the chance of future antisocial behavior by including the new comment for consideration: $p(\text{isUncivil}(c_{n+2})) = \text{CRAFT}(\{c_1, c_2, \ldots, c_n, c_{n+1}\})$.

[2]We use the publicly available ChangeMyView CRAFT model from the ConvoKit package (https://convokit.cornell.edu).

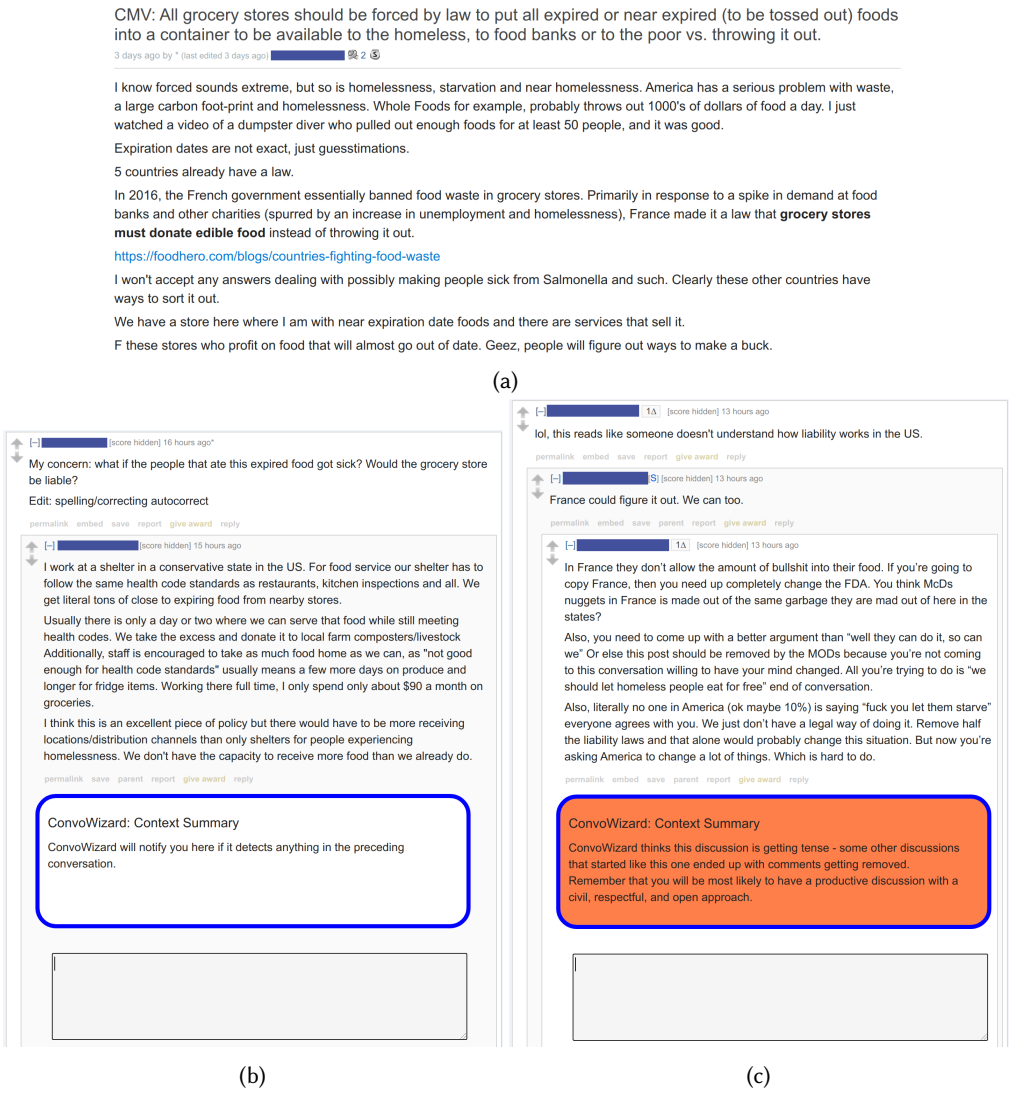

Fig. 1. The Context Summary feature of ConvoWizard provides information about whether the conversation the user is joining is at risk of turning uncivil in the future. (b) When no risk is detected, the Context Summary displays a neutral message on a blank background. (c) When risk is detected, the Context 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).

3.1.2 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 DOM,[3] 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

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

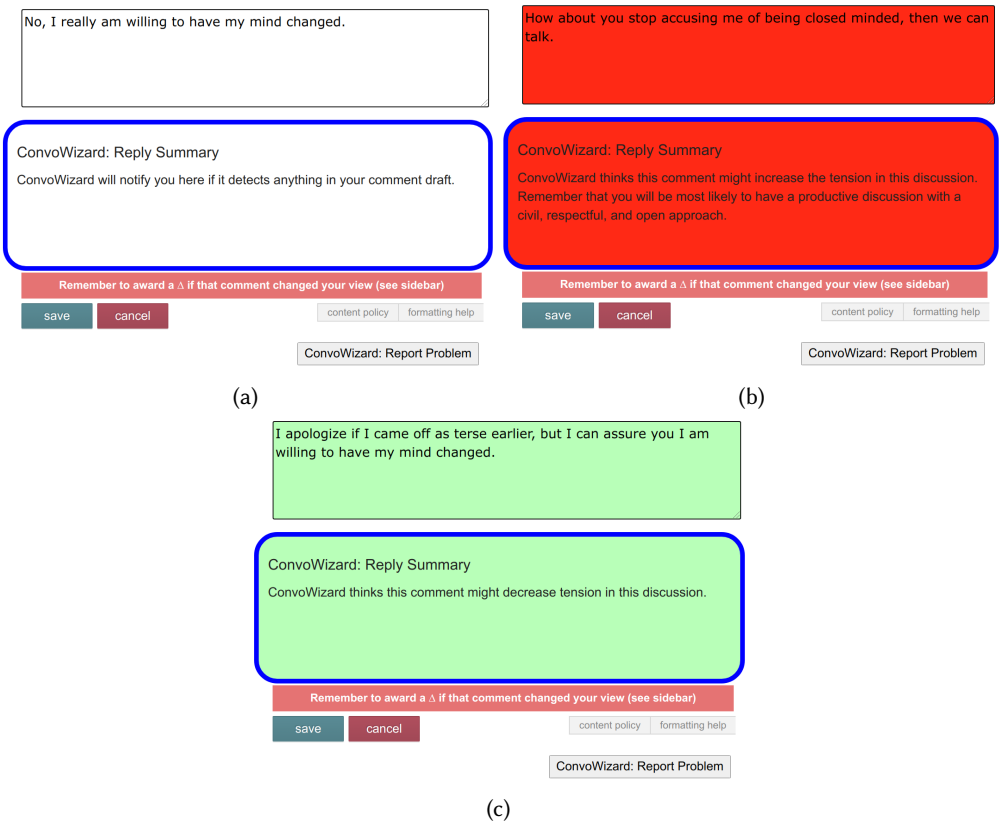

Fig. 2. The Reply Summary provides information about what impact the user's in-progress draft reply might have on the risk of incivility. (a) If the risk score with the draft reply is the same as the risk score without the draft reply (within a margin of error), the Reply Summary displays a neutral message. (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 1c; the preceding context is excluded for readability.)

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 3.1.3) 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 $S_{context} > 0.55$,[4] the Context Summary displays a warning to the user that the conversation they are about to participate in is

[4]The 0.55 threshold is what was recommended in the original CRAFT paper.

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 1 and in the Video Figure.

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 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 $S_{r(t)}$.[5] 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 $S_{context} - S_{r(t)}$.[6] This functionality is visualized in Figure 2 and in the Video Figure.

3.1.3 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 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 4.

3.1.4 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 3.2.1, 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" [53] or an outsider seeking to "brigade" the subreddit [26]).

[5] All comparisons apply a small noise threshold to prevent basing feedback on spurious variance in scores.
[6] 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.

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 [23, 29] and encode biases present in their training data [20]. But in the public consciousness, the capabilities and objectivity of algorithms are often overestimated [5, 46]. 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 1 and 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 continue to discuss in Section 3.2.1.

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.

3.2.1 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 3.1.4) 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 [37, 45, 73, 76],[7] 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

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

build solidarity and construct shared norms [22, 27, 67] and even serve as their community's representatives to the outside world [69].[8] 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,[9] 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 and updates which they could add to the pinned post.

3.2.2 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 in Section 3.2.3).

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.

[8] 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).

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

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 4). We return to discuss this limitation—and the steps needed for future work to overcome it—in more detail in Section 5.

3.2.3 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 A.

The exit survey can be roughly divided into three sections, all of which contain a mix of multiple-choice questions and open-ended text responses. First, we asked about participants' prior experiences with incivility and moderation, including what effects they think incivility has on discussions, how they personally react to incivility, and how effective moderation has been in their experience. Next, we asked participants to describe how they 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. Finally, we asked participants to give their general 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.

3.2.4 Data collection and processing. As described in Section 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 different 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).[10] Any interactions that could not be matched are discarded. This procedure results in a total of 334 pairs (668 total interactions).

[10]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.

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 both report a general willingness to take steps to mitigate risk of derailment, and identify key ways in which ConvoWizard's algorithmically-provided risk awareness helps them in that process. 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 rest of this section is organized as follows:

(1) First, we ask whether, independent of any external assistance, well-intentioned users are already willing to take steps to reduce tension in conversations that they feel are at risk, and if so, what concrete strategies they engage in. Through exit survey responses, we find that most users do self-report that they act proactively to reduce tension, and their responses give further insights into the specific proactive strategies they employ (Section 4.1).
(2) Next, we ask 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 4.2).
(3) Finally, in light of the observation that users are willing to act proactively and find algorithmic input helpful in deciding when to do so, 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 4.3).

4.1 Users' Intuitive Perception and Management of Risk

Responses from the exit survey indicate that in their regular commenting behavior (that is, in the absence of any outside assistance), well-intentioned users already proactively engage in some strategies for reducing tension when they intuitively deem it necessary. First, every participant reported that they have some level of intuition for when a discussion is at risk of turning uncivil. Explanations of how this intuition works vary across participants. Some participants reason about risk in terms of specific word choices:

> **P7**: 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".

> **P17**: 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:

> **P34**: 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.
>
> **P33**: 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.
>
> **P18**: [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 incivility, and 76.6% say that if they do join they will change how they phrase their reply. The latter finding in particular, that participants change their phrasing in response to perceived risk, is a promising indicator that they are willing to put in effort to reduce, or at least avoid increasing, the tension in the conversation. We explore this possibility further by examing what specific language changes these participants report. To focus the scope of this analysis, 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 [8, 55], and recent work has empirically validated this [81].
- **Formality**: Formality has been theorized to play a role in preventing misunderstanding [36], and in turn misunderstanding has been identified as a potential driver of incivility [13].
- **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 [34].
- **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 [39].
- **Swearing**: Swearing can be used to express aggression, but also to signal group identity or informality [38].
- **Comment length**: In the context of debates, higher word count can indicate that the interlocutor is trying to be more explicit in their argument [61, 62], which may, like formality, reflect an attempt to avoid misunderstanding.

The exit 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%).[11]

Overall, these findings show that well-intentioned users desire to avoid escalating at-risk conversations, and are willing to alter their behavior in order to achieve this goal. However, this does

[11]We also offered a free response "Other" option so participants could describe strategies that don't fit under any of the listed options. A quarter of the participants took this option, and a random sample of their responses can be found in Appendix B.

not imply that they are immune to engaging in uncivil 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 uncivil. These regrettable actions may be driven by a number of factors. For one, sometimes users may be making an inaccurate judgment of risk; as **P39** puts it:

> **P39**: 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 **P44** explains:

> **P44**: 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 **P39** and **P44**'s sentiments. Furthermore, even if the user does make an accurate judgment, they may still simply get caught up in the heat of the moment [15].

In light of this, we speculate that an additional nudge that enhances a user's awareness of existing tension in the conversation might prevent them from escalating the tension or even from crossing the line into uncivil behavior. This is a potential opening for algorithmic risk awareness interventions, and in the following sections we proceed to investigate this possibility.

4.2 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).[12]

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%).[13]

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 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. **P18** explains:

> **P18**: 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.

[12] 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 A.

[13] 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.

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 **P13** found:

> **P13**: 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. **P15** elaborates:

> **P15**: 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 **P2** explains:

> **P2**: 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 **P37** succinctly puts it:

> **P37**: 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 **P11** explains, it can leave users knowing that a conversation is at risk but not knowing what to do about it:

> **P11**: 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 **P13** identifies:

> **P13**: 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 [79]) 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.

4.3 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 prosocial 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 [52], and making (token-level) adjustments to their language use [68]. 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 3.2.4).

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.

4.3.1 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.'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" [52]—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, **P26** points out how seeing a warning from ConvoWizard might prompt them to review the conversational context more deeply than they would otherwise:

		Drafting time (seconds)
At-risk	Control	174.2
	Treatment	**189.5**
Not-at-risk	Control	124.3
	Treatment	133.4

(a)

		Correlation between adjusted timestamp and risk score
At-risk	Control	0.05^{**}
	Treatment	-0.06^{***}
Not-at-risk	Control	-0.13^{***}
	Treatment	-0.06^{**}

(b)

Table 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 ($^{**}p < 0.01$,$^{***}p < 0.001$).

> **P26**: 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.

P9 notes that even though they were aware the algorithm is imperfect, it was good enough to prompt reflection on their own in-progress draft:

> **P9**: I'm sure its not perfect, but in my case it made me rethink what I type.

Finally, **P2** explicitly mentions spending extra time rewording their comments:

> **P2**: 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 **P2** 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 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 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):

		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 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.

> **P35**: I kept rewording my reply until it stopped showing up orange.
>
> **P9**: 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 **P35** and **P9**, 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 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.[14] 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 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.

4.3.2 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 **P9** and **P18** explain:

[14] 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).

> **P9**: I thought of better words I could use maybe words that don't sound like I may be trying to provoke a uncivil response.
>
> **P18**: 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 that they had previously identified as their reactions to intuitively perceived risk: 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. 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 [68]. Our specific choice of summary variables is inspired by—but not exhaustive of[15]—the strategies that users self-reported employing in order to reduce tension:

- **F-factor**: This is a simple measure of *formality* introduced in [36]. 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 the authors of the F-factor 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 [63]. We use this score as it roughly corresponds to our definition of the *objective-subjective* distinction. It is computed as:

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

  We note that the CDI is one of the metrics used for quantifying the effects of prosocial interventions in [68].
- **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 [82], and so we adopt this heuristic.

[15]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 [19]).

Table 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 a number of notable differences in the comparisons. Compared to users in Control, users in Treatment ask more questions and are more likely to write comments that are judged as "formal" (according to the discretized F-factor);[16] both these differences are significant at $p < 0.05$. Furthermore, like the drafting effects, these effects are only found in at-risk interactions,[17] suggesting that, as expected, they are specific reactions to warnings. Finally, we find a similar trend (bordering on significance) in the CDI scores, with users in Treatment writing comments with a higher CDI; this is again specific to at-risk interactions.

As was the case with earlier analyses, these differences, while significant, 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, participant responses previously quoted in Section 4.1 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 4.2) 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 DISCUSSION

This work starts from the viewpoint that the solution to incivility in online discussions should come, in part, from the participants in these discussions. They can—and, as they indicate in our exit survey, 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, we propose a new proactive paradigm which seeks to prompt participants to employ their prosocial conversational skills by enhancing their awareness about the risks

[16] We note that the raw F-factor comparison is not significant, potentially due to the noisiness described before.
[17] The apparent increase in questioning in not-at-risk interactions as well is not statistically significant.

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 [12], and there can be a fine line between incivility and mere disagreement [1]. 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 [30] 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 [32]. 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 [18], 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 3.1.4) and running our study as a collaborative effort with community input (Section 3.2.1). 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 [79].

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 community 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 [41]. A public deployment of a tool like ConvoWizard would likewise be vulnerable to misuse; for example, as described in Section 3.1.4, 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 [56, 59] and flagging/reporting systems [18] 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 [65]. 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 [67]. 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 [42]. 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 downstream effects, with a particular eye on whether the prosocial changes triggered by a ConvoWizard warning (Section 4.3.1) might further translate to more civil behavior of other interlocutors [3], or whether they strengthen or weaken the persuasive effectiveness of the argument [73]. 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 4.2).

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 [73]. 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 [59] as well as work-coordination settings like Wikipedia Talk Pages [48, 80], have an added incentive to keep discussions civil since incivility can distract from their broader non-conversational goals [1]. 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.

ACKNOWLEDGEMENTS

We would like to thank Michael Bernstein, Lillian Lee, Karen Levy, Cecelia Madsen, the 2021-2022 cohort of fellows at the Center for Advanced Study in the Behavioral Sciences at Stanford, and all the reviewers for the enlightening discussions and helpful suggestions. We additionally recognize everyone who helped with the implementation of ConvoWizard, particularly Lucas Van Bramer and Oscar So for their contributions to the codebase, Todd Cullen for his help in setting up the backend server configuration, and Caleb Chiam, Liye Fu, Khonzoda Umarova, and Justine Zhang for their extensive testing and generous feedback. Finally, we are grateful to all the ChangeMyView users who participated in our study, and to the ChangeMyView moderators for their guidance and for serving as a point of contact to the broader ChangeMyView community. This research was supported in part by an NSF CAREER award IIS-1750615 and by an NSF Grant IIS-1910147; Jonathan P. Chang was supported in part by a fellowship with the Cornell Center for Social Sciences and Cristian Danescu-Niculescu-Mizil was supported in part by fellowships with the Cornell Center for Social Sciences and with the Center for Advanced Study in the Behavioral Sciences at Stanford.

REFERENCES

[1] Ofer Arazy, Lisa Yeo, and Oded Nov. 2013. Stay on the Wikipedia Task: When Task-related Disagreements Slip Into Personal and Procedural Conflicts. *J. Assoc. Inf. Sci. Technol.* 64, 8 (Aug. 2013).
[2] Zahra Ashktorab and Jessica Vitak. 2016. Designing Cyberbullying Mitigation and Prevention Solutions through Participatory Design With Teenagers. In *Proceedings of CHI*.
[3] Jiajun Bao, Junjie Wu, Yiming Zhang, Eshwar Chandrasekharan, and David Jurgens. 2021. Conversations Gone Alright: Quantifying and Predicting Prosocial Outcomes in Online Conversations. In *Proceedings of WWW*.
[4] Matt Billings and Leon A. Watts. 2010. Understanding Dispute Resolution Online: Using Text to Reflect Personal and Substantive Issues in Conflict. In *Proceedings of CHI*.
[5] Paolo Bory. 2019. Deep New: The Shifting Narratives of Artificial Intelligence from Deep Blue to AlphaGo. *Convergence* 25, 4 (Aug. 2019).
[6] Johanna Brewer, Morgan Romine, and T. L. Taylor. 2020. Inclusion at Scale: Deploying a Community-Driven Moderation Intervention on Twitch. In *Proceedings of DIS*.
[7] Barry Brown, Stuart Reeves, and Scott Sherwood. 2011. Into the Wild: Challenges and Opportunities for Field Trial Methods. In *Proceedings of CHI*.
[8] Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage.* Cambridge University Press.
[9] Jie Cai and Donghee Yvette Wohn. 2019. What Are Effective Strategies of Handling Harassment on Twitch? Users' Perspectives. In *Proceedings of CSCW*.
[10] Jie Cai, Donghee Yvette Wohn, and Mashael Almoqbel. 2021. Moderation Visibility: Mapping the Strategies of Volunteer Moderators in Live Streaming Micro Communities. In *Proceedings of IMX*.
[11] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. In *Proceedings of CSCW*.
[12] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. In *Proceedings of CSCW*.
[13] Jonathan P Chang, Justin Cheng, and Cristian Danescu-Niculescu-Mizil. 2020. Don't Let Me Be Misunderstood:Comparing Intentions and Perceptions in Online Discussions. In *Proceedings of WWW*.

[14] Jonathan P. Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trouble on the Horizon: Forecasting the Derailment of Online Conversations as They Develop. In *Proceedings of EMNLP*.
[15] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In *Proceedings of CSCW*.
[16] Emily I. M. Collins, Anna L. Cox, Jon Bird, and Daniel Harrison. 2014. Social Networking Use and RescueTime: The Issue of Engagement. In *Proceedings of UbiComp Adjunct*.
[17] 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. 2008. Activity Sensing in the Wild: A Field Trial of Ubifit Garden. In *Proceedings of CHI*.
[18] Kate Crawford and Tarleton Gillespie. 2016. What Is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint. *New Media & Society* 18, 3 (March 2016).
[19] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A Computational Approach to Politeness with Application to Social Factors. In *Proceedings of ACL*.
[20] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of ICWSM*.
[21] Julian Dibbell. 2005. A Rape in Cyberspace. *The Village Voice* (Oct. 2005).
[22] Bryan Dosono and Bryan Semaan. 2019. Moderation Practices as Emotional Labor in Sustaining Online Communities: The Case of AAPI Identity Work on Reddit. In *Proceedings of CHI*.
[23] Natasha Duarte and Emma Llansó. 2018. Mixed Messages? The Limits of Automated Social Media Content Analysis. In *Proceedings of FAccT*.
[24] R. Stuart Geiger. 2016. Bot-Based Collective Blocklists in Twitter: The Counterpublic Moderation of Harassment in a Networked Public Space. *Information, Communication & Society* 19, 6 (June 2016).
[25] R. Stuart Geiger and David Ribes. 2010. The Work of Sustaining Order in Wikipedia: The Banning of a Vandal. In *Proceedings of CSCW*.
[26] Alex Georgakopoulou, Stefan Iversen, and Carsten Stage. 2020. Making Memes Count: Platformed Rallying on Reddit. In *Quantified Storytelling: A Narrative Analysis of Metrics on Social Media*, Alex Georgakopoulou, Stefan Iversen, and Carsten Stage (Eds.). Springer International Publishing, Cham.
[27] Sarah A. Gilbert. 2020. "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*.
[28] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press, New Haven.
[29] Tarleton Gillespie. 2020. Content Moderation, AI, and the Question of Scale:. *Big Data & Society* (July 2020).
[30] Tarleton Gillespie, Patricia Aufderheide, Elinor Carmi, Ysabel Gerrard, Robert Gorwa, Ariadna Matamoros-Fernández, Sarah T. Roberts, Aram Sinnreich, and Sarah Myers West. 2020. Expanding the Debate about Content Moderation: Scholarly Research Agendas for the Coming Policy Debates. *Internet Policy Review* 9, 4 (Oct. 2020).
[31] Joseph K. Goodman, Cynthia E. Cryder, and Amar Cheema. 2013. Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples. *Journal of Behavioral Decision Making* 26, 3 (2013).
[32] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance. *Big Data & Society* 7, 1 (Jan. 2020).
[33] James Grimmelmann. 2015. The Virtues of Moderation. *Yale Journal of Law and Technology* 17, 1 (Sept. 2015).
[34] Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. Before Name-Calling: Dynamics and Triggers of Ad Hominem Fallacies in Web Argumentation. In *Proceedings of NAACL*.
[35] Aaron Halfaker, Bryan Song, D. Alex Stuart, Aniket Kittur, and John Riedl. 2011. NICE: Social Translucence Through UI Intervention. In *Proceedings of WikiSym*.
[36] Francis Heylighen and Jean-Marc Dewaele. 1999. *Formality of Language: Definition, Measurement and Behavioral Determinants*. Technical Report. Center "Leo Apostel", Free University of Brussels.
[37] Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the Semantic Types of Claims and Premises in an Online Persuasive Forum. In *Proceedings of the 4th Workshop on Argument Mining*.
[38] Eric Holgate, Isabel Cachola, Daniel Preotiuc-Pietro, and Junyi Jessy Li. 2018. Why Swear? Analyzing and Inferring the Intentions of Vulgar Expressions. In *Proceeding of EMNLP*.
[39] Karen Huang, Michael Yeomans, Alison Wood Brooks, Julia Minson, and Francesca Gino. 2017. It Doesn't Hurt to Ask: Question-asking Increases Liking. *Journal of Personality and Social Psychology* 113, 3 (Sept. 2017).
[40] Krithika Jagannath, Katie Salen, and Petr Slovàk. 2020. "(We) Can Talk It Out...": Designing for Promoting Conflict-Resolution Skills in Youth on a Moderated Minecraft Server. In *Proceedings of CSCW*.
[41] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. "Did You Suspect the Post Would Be Removed?": Understanding User Reactions to Content Removals on Reddit. In *Proceedings of CSCW*.

[42] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Transactions on Computer-Human Interaction* 26, 5 (July 2019).
[43] Shagun Jhaver, Larry Chan, and Amy Bruckman. 2018. The View from the Other Side: The Border between Controversial Speech and Harassment on Kotaku in Action. *First Monday* 23, 2 (Feb. 2018).
[44] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online Harassment and Content Moderation: The Case of Blocklists. *ACM Transactions on Computer-Human Interaction* 25, 2 (March 2018).
[45] Shagun Jhaver, Pranil Vora, and Amy Bruckman. 2017. *Designing for Civil Conversations: Lessons Learned from ChangeMyView.* Technical Report. Georgia Institute of Technology.
[46] Christian Katzenbach. 2021. "AI Will Fix This" – The Technical, Discursive, and Political Turn to AI in Governing Communication. *Big Data & Society* 8, 2 (July 2021).
[47] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. 2012. Regulating Behavior in Online Communities. In *Building Successful Online Communities: Evidence-Based Social Design*, Paul Resnick and Robert Kraut (Eds.). MIT Press.
[48] Aniket Kittur and Robert E. Kraut. 2008. Harnessing the Wisdom of Crowds in Wikipedia: Quality Through Coordination. In *Proceedings of CSCW.*
[49] Geza Kovacs, Zhengxuan Wu, and Michael S. Bernstein. 2018. Rotating Online Behavior Change Interventions Increases Effectiveness But Also Increases Attrition. In *Proceedings of CSCW.*
[50] Paul Krebs, James O. Prochaska, and Joseph S. Rossi. 2010. A Meta-Analysis of Computer-Tailored Interventions for Health Behavior Change. *Preventive Medicine* 51, 3 (Sept. 2010).
[51] Travis Kriplean, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett. 2012. Supporting Reflective Public Thought with Considerit. In *Proceedings of CSCW.*
[52] Travis Kriplean, Michael Toomim, Jonathan Morgan, Alan Borning, and Andrew Ko. 2012. Is This What You Meant?: Promoting Listening on the Web with Reflect. In *Proceedings of CHI.*
[53] Srijan Kumar, Justin Cheng, Jure Leskovec, and V.S. Subrahmanian. 2017. An Army of Me: Sockpuppets in Online Discussion Communities. In *Proceedings of WWW.*
[54] Srijan Kumar, William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Community Interaction and Conflict on the Web. In *Proceedings of WWW.*
[55] Robin T. Lakoff. 1973. *The Logic of Politeness: Minding Your P's and Q's.* Chicago Linguistic Society.
[56] Cliff Lampe and Paul Resnick. 2004. Slash(Dot) and Burn: Distributed Moderation in a Large Online Conversation Space. In *Proceedings of CHI.*
[57] Ping Liu, Joshua Guberman, Libby Hemphill, and Aron Culotta. 2018. Forecasting the Presence and Intensity of Hostility on Instagram Using Linguistic and Social Features. In *Proceedings of ICWSM.*
[58] Claudia (Claudia Wai Yu) Lo. 2018. *When All You Have Is a Banhammer : The Social and Communicative Work of Volunteer Moderators.* Thesis. Massachusetts Institute of Technology.
[59] Lena Mamykina, Bella Manoim, Manas Mittal, George Hripcsak, and Björn Hartmann. 2011. Design Lessons from the Fastest Q&A Site in the West. In *Proceedings of CHI.*
[60] Aske Mottelson and Kasper Hornbæk. 2017. Virtual Reality Studies Outside the Laboratory. In *Proceedings of VRST.*
[61] Daniel J. O'Keefe. 1997. 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).
[62] Daniel J. O'Keefe. 1998. Justification Explicitness and Persuasive Effect: A Meta-Analytic Review of the Effects of Varying Support Articulation in Persuasive Messages. *Argumentation and Advocacy* 35, 2 (Sept. 1998).
[63] James W. Pennebaker, Cindy K. Chung, Joey Frazee, Gary M. Lavergne, and David I. Beaver. 2014. When Small Words Foretell Academic Success: The Case of College Admissions Essays. *PLOS ONE* (2014).
[64] Katharina Reinecke and Krzysztof Z. Gajos. 2015. LabintheWild: Conducting Large-Scale Online Experiments With Uncompensated Samples. In *Proceedings of CSCW.*
[65] Annika Richterich. 2014. 'Karma, Precious Karma!' Karmawhoring on Reddit and the Front Page's Econometrisation. *Journal of Peer Production* 4, 1 (2014).
[66] Sarah T Roberts. 2014. *Behind the Screen: The Hidden Digital Labor of Commercial Content Moderation.* Ph. D. Dissertation. University of Illinois at Urbana-Champagne.
[67] Joseph Seering. 2020. Reconsidering Self-Moderation: The Role of Research in Supporting Community-Based Models for Online Content Moderation. In *Proceedings of CSCW.*
[68] Joseph Seering, Tianmi Fang, Luca Damasco, Mianhong 'Cherie' Chen, Likang Sun, and Geoff Kaufman. 2019. Designing User Interface Elements to Improve the Quality and Civility of Discourse in Online Commenting Behaviors. In *Proceedings of CHI.*
[69] Joseph Seering, Geoff Kaufman, and Stevie Chancellor. 2020. Metaphors in Moderation. *New Media & Society* (Oct. 2020).
[70] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of CSCW.*

[71] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator Engagement and Community Development in the Age of Algorithms. *New Media & Society* 21, 7 (July 2019).
[72] Kumar Srinivasan, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan. 2019. Content Removal as a Moderation Strategy: Compliance and Other Outcomes in the ChangeMyView Community. In *Proceedings of CSCW*.
[73] Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu, and Lillian Lee. 2016. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions. In *Proceedings of WWW*.
[74] Samuel Hardman Taylor, Dominic DiFranzo, Yoon Hyung Choi, Shruti Sannon, and Natalya N. Bazarova. 2019. Accountability and Empathy by Design: Encouraging Bystander Intervention to Cyberbullying on Social Media. In *Proceedings of CSCW*.
[75] Kal Turnbull. 2018. "That's Bullshit" – Rude Enough for Removal? A Multi-Mod Perspective.
[76] Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is This Post Persuasive? Ranking Argumentative Comments in the Online Forum. In *Proceedings of ACL*.
[77] Galen Weld, Amy X. Zhang, and Tim Althoff. 2022. What Makes Online Communities 'Better'? Measuring Values, Consensus, and Conflict across Thousands of Subreddits. In *Proceedings of ICWSM*.
[78] Donghee Yvette Wohn. 2019. Volunteer Moderators in Twitch Micro Communities: How They Get Involved, the Roles They Play, and the Emotional Labor They Experience. In *Proceedings of CHI*.
[79] Austin P. Wright, Omar Shaikh, Haekyu Park, Will Epperson, Muhammed Ahmed, Stephane Pinel, Diyi Yang, and Duen Horng Chau. 2021. RECAST: Interactive Auditing of Automatic Toxicity Detection Models. In *Proceedings of CSCW*.
[80] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of WWW*.
[81] Justine Zhang, Jonathan P. Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Nithum Thain, Yiqing Hua, and Dario Taraborelli. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of ACL*.
[82] Justine Zhang, Arthur Spirling, and Cristian Danescu-Niculescu-Mizil. 2017. Asking Too Much? The Rhetorical Role of Questions in Political Discourse. In *Proceedings of EMNLP*.

A USER STUDY AND EXIT SURVEY DETAILS

A.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.[18] 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 was offered). All participants who accepted the invitation to join Phase 2 chose the 60-day option.

A.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

[18] 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.

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.

A.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 participation (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 Appendix B 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/changemyview 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 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

Received January 2022; revised April 2022; accepted August 2022