AI Within Online Discussions: Rational, Civil, Privileged? Ethical Considerations on the Interference of AI in Online Discourse

Jonas Aaron Carstens & Dennis Friess

Suggested Citation:
Carstens, J. A., & Frieß, D. (2024). AI Within Online Discussions: Rational, Civil, Privileged?: Ethical Considerations on the Interference of AI in Online Discourse. Minds and Machines, 34(2), Article 10. https://doi.org/10.1007/s11023-024-09658-0

Minds and Machines (2024) 34:10
https://doi.org/10.1007/s11023-024-09658-0

AI Within Online Discussions: Rational, Civil, Privileged?

Ethical Considerations on the Interference of AI in Online Discourse

Jonas Aaron Carstens[1] · **Dennis Friess**[2]

Received: 1 June 2023 / Accepted: 21 January 2024 / Published online: 4 May 2024

Abstract
While early optimists have seen online discussions as potential spaces for deliberation, the reality of many online spaces is characterized by incivility and irrationality. Increasingly, AI tools are considered as a solution to foster deliberative discourse. Against the backdrop of previous research, we show that AI tools for online discussions heavily focus on the deliberative norms of rationality and civility. In the operationalization of those norms for AI tools, the complex deliberative dimensions are simplified, and the focus lies on the detection of argumentative structures in argument mining or verbal markers of supposedly uncivil comments. If the fairness of such tools is considered, the focus lies on data bias and an input–output frame of the problem. We argue that looking beyond bias and analyzing such applications through a sociotechnical frame reveals how they interact with social hierarchies and inequalities, reproducing patterns of exclusion. The current focus on verbal markers of incivility and argument mining risks excluding minority voices and privileges those who have more access to education. Finally, we present a normative argument why examining AI tools for online discourses through a sociotechnical frame is ethically preferable, as ignoring the predicable negative effects we describe would present a form of objectionable indifference.

Keywords AI · Discourse · Deliberation · Fairness · Equality · Discrimination

✉ Jonas Aaron Carstens
Jonas.Carstens@hhu.de

[1] Department of Political Philosophy and Ethics, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

[2] Düsseldorf Instute for Internet and Democracy, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

1 Introduction

The advent of the internet has led to the emergence of multiple online publics, where people from many different backgrounds are able to discuss issues of public manner. While some authors have argued that the internet could provide the infrastructure for a more deliberative public sphere (e.g., Coleman & Gøtze, 2001; Dahlberg, 2001), others have raised concerns that online communication could reveal the darkest human abysses (Papacharissi, 2004; Suler, 2004). More than two decades later, one may argue that these pessimistic assessments are empirically evident. In fact, several studies indicate that online discussions suffer in terms of civility and are far away from reasoned and democratic discourse envisioned by the advocates of deliberative democracy (Coe et al., 2014; Kreissel et al., 2018). Against this background, Artificial Intelligence (AI) has entered this research field in recent years, which means that both scholars and commercial organizations develop and employ AI-driven tools in order to maintain democratic discussions online (Rodríguez-Ruiz et al., 2020; Stoll et al., 2020; Wojcieszak et al., 2021). In this paper, we focus on AI that aims to improve the quality of online discussions.

Following Hancock et al., (2020, p. 90), AI broadly refers "to computational systems that involve algorithms, machine learning methods, natural language processing [NLP], and other techniques that operate on behalf of an individual to improve a communication outcome". However, AI aiming to improve communication quality may also trigger ethical issues, which constitute our main point of interest. While scholars and developers may intend to improve the quality of public online discourses when introducing automated hate speech detection tools, argument mining models, and moderating bots, they may also unintentionally increase existing inequalities, exclude certain voices from the discourse, and deepen social hierarchies.[1] Since determining what is a solid argument, an appropriate wording or a comment worth to be automatically replied to by a bot has powerful implications; such decisions have to be the subject of ethical reflections. Those reflections are in place for some AI applications—e.g. medical diagnostic tools or credit scoring, but are less developed in the context of AI interfering in public online discourses.

In the first part of the paper, we are going to sketch the state of online discourse, arguing that norms of deliberation are still considered to be important normative standards to evaluate the quality of online discussions and that the violation of these norms has paved the way for AI to clean up discussions (1). In the next step, we will provide some orientation on previous AI research in the context of online deliberation to illustrate that AI research already aims to improve certain norms of deliberation while neglecting others (2). Zooming in on rationality and civility, we discuss how those norms are conceptualized and subsequently operationalized for AI tools.

[1] We use the term social hierarchies to refer to systematic differences in the power and authority that members of different groups hold within society. Power is understood as the amount of possible actions available as well as the potential to compel others to action, whereas authority refers to the ability to be recognized and listened to as well as the potential to be ascribed expertise (see for power and authority Moreau 2020, pp. 51–52). Inequality, on the other hand, refers to broader patterns of different levels of access to goods and opportunities between groups.

Furthermore, we discuss how such AI tools are imagined as neutral once the data bias that influences the pipeline from input to output has been addressed (3). We proceed by adopting a sociotechnical perspective, moving beyond data bias, showing that the promotion of deliberative norms through AI can exclude and silence marginalized groups by ignoring the interaction of models with cultural norms and social hierarchies. (4). Finally, we argue that restricting the analysis to the relationship between input and output constitutes an arbitrary choice, which expresses *objectionable indifference* towards those that are further excluded and marginalized through AI tools. To avoid expressing objectionable indifference, AI tools should be evaluated through a sociotechnical framework that explicitly addresses equality instead of presupposing neutrality (5). Finally, we provide concluding remarks (6).

2 The State of Online Discourse

Some early writings have painted the internet as a virtual public space for free-flowing discussions and the respectful exchange of arguments (Dahlberg, 2001; Negroponte, 1995). Particularly, advocates of deliberative democracy have argued that the internet would provide the communicative infrastructure for large-scale deliberation by a diverse and dispersed public without significant limitations with regard to time and space (Dahlberg, 2007; Graham & Witschge, 2003). Furthermore, online discussions and public participation platforms could foster civic engagement in political decision-making in a deliberative manner (Aitamurto & Landemore, 2013; Nelimarkka et al., 2014; Romberg & Conrad, 2021). Against this backdrop, Chadwick (2009, p. 13) stated that deliberation is probably the most influential concept in the context of digital democracy research. In fact, scholarship on online discussions often adopts a deliberative perspective by implicitly or explicitly drawing on norms that arise from theories of deliberative democracy (Esau et al., 2017; Ruiz et al., 2011).

In contrast, other authors have emphasized concerns and risks by arguing that instead of cultivating deliberative discussions, the online environment would transform people into digital rowdies, and abusive language, flaming, sexism, racism, and hate speech would spread online (Buchstein, 1996; Suler, 2004). In fact, empirical research suggests that online discussions often contain uncivil[2] contributions, such as hate speech, and therefore suffer in terms of deliberative quality (Coe et al., 2014). Research also suggests that such uncivil discourse is even orchestrated by organized networks (Garland et al., 2022; Kreissel et al., 2018). Such networks are also responsible for strategic misinformation campaigns (Pacheco et al., 2020), which are sometimes also supported by 'social' bots (Uyheng et al., 2022). These campaigns, where accusations of hoaxes and lies are frequent, curtail substantial engagement and the critical exchange of arguments and can lead to a situation where "communication has reached a dead end" (Brüggemann et al., 2020, p. 1026). Some online spaces

[2] Incivility here is preliminarily defined as "features of discussion that convey an unnecessarily disrespectful tone toward the discussion forum, its participants, or its topics" (Coe et al., 2014, p. 660).

are even dedicated to decisively irrational discourse, such as climate change denial, that often spills over to other sites (Davis, 2021). For example, some Twitter debates on climate change have been found to be characterized by affective polarization and hostility, especially on the side of deniers, while rational argument is lacking (Tyagi et al., 2020).

Research has shown that those low-quality discourses can influence readers' perceptions of public opinion toward specific issues (Anderson et al., 2014), their perceptions of news quality (Anderson et al., 2018), and their commenting behavior (Springer et al., 2015). In addition to that, incivility erodes the participants' expectations in the potential for success of public deliberation (Hsueh et al., 2015; Hwang et al., 2014). Furthermore, frequent exposure to incivility on social media can lead to its normalization and heighten negative and stereotypical perceptions of minorities (Soral et al., 2020).

The presence of the described phenomena and the findings on its negative effects have called both researchers and organizers of online public discourse to action in order to maintain democratic discourse (e.g., Friess et al., 2021; Stroud et al., 2015; Ziegele et al., 2018). However, since online spaces open to public discussions are not easily maintained and moderated by humans alone, research has addressed the question whether AI is able to solve some of the problems associated with online public discussions (e.g., Romberg & Conrad, 2021; Stoll et al., 2020). In the next section, we will focus on this strand of literature.

3 Previous Research on AI Promoting Deliberative Norms

Improving online discourse needs both a desirable ideal for what better discourse could look like and a method or tool for moving closer to that ideal. One influential expression of the ideal discourse we mentioned above is the concept of deliberation, which originates from the literature on deliberative democracy (see Friess & Eilders, 2015 for a review; Gutmann & Thompson, 2004; Habermas, 1998). In a nutshell, deliberation can be described as a demanding form of communication that is characterized by certain norms. Even though there are competing conceptions of which norms exactly constitute deliberation, many authors share the idea that deliberation is a constructive, reciprocal, and respectful exchange of reasons among equal participants (Bächtiger & Parkinson, 2019; Friess & Eilders, 2015). We will discuss this concept in more detail later on.

Recently, research has focused on AI as a tool to foster and maintain democratic discussions online that live up to standards of deliberation (Rodríguez-Ruiz et al., 2020; Stoll et al., 2020; Wojcieszak et al., 2021). In order to systematically access the potential of AI for deliberation, Friess et al. (2022) reviewed 171 scientific contributions from computational science and communication studies that have studied AI in the context of online deliberation. The review suggests that AI research has mostly focused on the dimensions of rationality and civility, while other norms of deliberation, such as reciprocity or equality, seem to be much less studied, yet Friess et al. (2022) show that a closer look at the dimension of rationality makes it apparent that *argument mining* constitutes a key focus when AI is used to enhance

rationality (e.g., Ida et al., 2019; Klein, 2022; Liebeck et al., 2016) while the automated detection of duplicates (Yang & Callan, 2006), topic modelling approaches (e.g., Curiskis et al., 2020), or stance detection (e.g., Sirrianni et al., 2021) are less frequently researched in the context of rationality. We will turn back to this evident focus on argument mining below.

With regard to the deliberative dimension of civility, studies on the automated detection of hate speech and offensive language are dominant in the reviewed literature (e.g., Daxenberger et al., 2018; Stoll et al., 2020). Beside the pure detection of incivility, some studies have also experimented with automated counterspeech bots (e.g., Clever et al., 2022). Even though different studies target different forms of deviant communication, this strand of literature has significantly advanced in the last 5 years and is already able to contribute to more civil discussions online.

Compared to rationality and civility, the deliberative dimensions of reciprocity and equality were much less targeted by the AI research, according to Friess et al. (2022). However, some studies, for example, have experimented with AI-based chat assistants to increase reciprocity among users (e.g., Argyle et al., 2023; Ito et al., 2022; Wyss & Beste, 2017). Furthermore, AI has been used to stimulate more balanced and equal participation in online discussions by automatically identifying the activity of users, using deep learning approaches (Wijenayake et al., 2020) or chatbots that encourage passive participants to actively contribute to the discussion (Kim et al., 2021). AI has also been used for automatic text simplification in order to increase inclusiveness of online discussions (Stodden et al., 2023).

However, the overwhelming prevalence of AI research that targets rationality and civility justifies a closer look at both the deliberative norms themselves and their implementation in AI tools. While we do not intend to disregard the importance of the dimensions of equality and reciprocity within the framework of deliberative theory (Friess & Eilders, 2015), the state of AI research seems to be skewed towards the theoretical dimensions of rationality and civility. Thus, we focus our ethical reflections on those norms. We will now discuss those two deliberative norms and their operationalization in the development of AI tools, before turning to the discussion of harmful effects of those tools.

4 Rationality and Civility: From Conceptualization to AI Tools

We will now examine more closely the conceptualization and the operationalization of rationality and civility. Afterwards, we will discuss how AI models used in tools to promote the two dimensions can potentially become biased and how the analysis of that bias often focuses on an input–output frame.

When engaging with the concept of rationality in online deliberation research, the work of Habermas provides a basis for much of the conceptualization (Dahlberg, 2010; Friess & Eilders, 2015). Habermas considers rationality constitutive for communicative practice and deems it central to rationality that a speaker who raises a claim can defend the claim with reasons (Habermas, 1995, p. 37). In case of dissent, rationality is tied to argumentation, as those exhibiting communicative rationality back their disputed claims with arguments or revise their utterance in response to

criticism (Habermas, 1995, p. 38). Such communication is best understood as a normative goal. Ideally, participants do not only factually reach consent but aim at motivating consent through critical reasoning (Habermas, 1995, p. 525). Under those conditions, communicative argumentation can be understood as the cooperative search for truth (Habermas, 1983, p. 98). Despite the common roots in Habermas' work, rationality as a central aspect of deliberative theory is not operationalized in a unified way across different works (Friess & Eilders, 2015). However, Habermas' reliance on arguments emerges again, for example, when Dahlberg (2010, p. 623) considers it central to deliberative discourse that "positions […] are provided with reasons rather than simply asserted". Dutwin (2003) identifies rationality with reasoned argument. Generally, argumentation plays a central role in operationalizing rationality (Friess & Eilders, 2015; Stromer-Galley, 2007). The focus on arguments supporting a claim also emerges as evident from the presented literature review, as most AI tools in the dimension of rationality focus on argument mining.

Argument mining denotes "the automatic identification and extraction of argument components and structure" (Lawrence & Reed, 2020, p. 766). However, since considerable dissent exists, "it is impossible to give a single formal, universally accepted definition of structured argument" (Lippi & Torroni, 2015, p. 165). Thus, the following steps can vary according to the underlying model of argumentation. Argument mining is based on *Natural Language Processing* (NLP) which provides the possibility to automate processing of naturally occurring texts (Cabrio & Villata, 2018). In the past, this has often involved machine learning algorithms trained on pre-labeled datasets in supervised learning conditions (Liu et al., 2023). However, the training of NLP models has shifted in recent cases to language models trained on raw data, not on prelabeled instances, providing larger datasets for training (Liu et al., 2023). Thus, argument mining models can be trained on manually annotated corpora (Lippi & Torroni, 2015; Ruckdeschel & Wiedemann, 2022), but pretrained language models can also be fine-tuned on unlabeled data (Dutta et al., 2022). Models for argument mining often involve the prediction or identification of argument boundaries to identify which parts of a text contain arguments (Fu et al., 2023; Lippi & Torroni, 2015). Input that features narratives, questions, or simple expression of agreement is then discarded and classified as non-argumentative (Habernal & Gurevych, 2017). Moreover, the components of an argument, such as claims and different aspects of argumentative units, can be determined (Lippi & Torroni, 2015; Ruckdeschel & Wiedemann, 2022). Finally, the intercomponent relationship between the different components of an argument can be predicted (Dutta et al., 2022).

Applications of argument mining, for example, take the form of automating the extraction of arguments in online participation processes and online discussions as well as identifying trends and contentious issues (Lawrence et al., 2017; Liebeck et al., 2016). Furthermore, argument mining can be used to predict parts of an argument that are most susceptible to attacks (Jo et al., 2020), which could be used to facilitate more engagement from opposing sides (Vecchi et al., 2021). Finally, argument mining can potentially be used in collaborative online discussions, supporting users by summarizing the state of a debate or even by providing feedback on their own arguments so they can "be nudged into writing more persuasive arguments"

(Schneider, 2014, p. 61). Ida et al. (2019) present a concept of a discussion-inducing forum that uses argument mining to present the argument structures of contributions to participants. Thus, argument mining can be used to assess the results of a debate for both users and decision-makers, but also to actively influence a debate.

From the Habermasian idea of rationality to the operationalization and finally the use in AI tools, a strong focus on formal argument structures emerges. Hence, the complex conceptualization that Habermas presents is mostly reduced to a simpler operationalization. Aspects such as cooperative truth-seeking are omitted or supposed to be promoted by the identification of specific argumentative structures.

Next to rationality, civility emerges as a particularly important deliberative dimension in the context of AI tools for improving online discourse. Whereas rationality is a mainstay of Habermas' work, civility is less directly addressed but can be indirectly reconstructed, such as from the idea that the discourse must be free from any external and implicit pressure or repression (Habermas, 1983, p. 99). Furthermore, Habermas understands the ideal discourse as a fundamentally cooperative process of mutual understanding and truth-seeking in which the other can be supposed to act truthfully and honestly (Habermas, 1983, pp. 97–98). Much of the disrespectful behavior described above, such as intimidation and disparaging others, can be understood as implicit pressure or repression endangering any cooperative process.

In the literature on online deliberation, several competing definitions of civility and incivility exist (Bormann & Ziegele, 2023). Friess and Eilders (2015, p. 330) understand civility as "mutual recognition of the participants in the sense that everybody is recognized as an equal actor able to speak in his or her own manner". Additionally, respect can be considered as a central aspect of civility. Coe et al., (2014, p. 660) define incivility as contributions that are unnecessarily disrespectful in tone, emphasizing that such content does not add "anything of substance". When measuring incivility and disrespect, Coe et al. (2014) include speech acts that are disparaging or vulgar. Papacharissi (2004, p. 267) in contrast, cautions that simple rudeness should not be considered uncivil because such a definition may stifle heated debate. She argues that the focus should lie on behaviors that risk setting back a democratic society especially by attacking social groups of which a discussion participant is a member (2004, p. 267). All in all, those differing understandings provide plausibility to the judgement of Bormann and Ziegele (2023, p. 211) that incivility will likely be "subject to individual perceptions and zeitgeist".

When it comes to developing AI tools, the complexity of the debate is mostly foregone for the detection of vulgarity or disparaging remarks (Davidson et al., 2020; Sadeque et al., 2019; Vidgen & Derczynski, 2020). Sadeque et al. (2019) present a Neural Network Model that can detect and filter name-calling and vulgarity, trained on annotated data. Relying on pre-trained models, Davidson et al. (2020) present a model that also includes disparaging communication towards an idea or disparaging remarks about the way a person communicates, besides name-calling and vulgarity. Vidgen and Derczynski (2020) discuss in their review article of abusive language detection that incivility detection often relies on identifying the tone of the contribution. The reliance on aspects such as name-calling and vulgarity makes it necessary to at least partly rely on linguistic markers in the annotation

process and subsequently in the use of AI tools. Once again, from the concept of civility to its operationalization in AI tools, a significant reduction of complexity appears. While the concept remains subject of debate in deliberative research, AI tools rely on simpler markers like name-calling or vulgarity, ignoring more nuanced definitions of respect and the multitude of possible operationalizations (Bormann & Ziegele, 2023, p. 201).

In the use of AI, data bias is a central concern (Ntoutsi et al., 2020), and tools for online discussions are no exception. Concerns about data bias mainly aim at identifying ways in which a model can acquire biases and consequently render the process from the input data to the output of the model unfair. This debate about the neutrality and fairness[3] of algorithms is decades old (see for an early example Lowry & Macpherson, 1988). Since then, the mechanisms through which bias can enter algorithmic decision-making have been extensively studied and described, with a strong focus on machine learning (Barocas & Selbst, 2016; Jiang & Nachum, 2020). Those discoveries have contributed to a growing awareness that the incorporation of bias is a problem for AI applications and spawned a variety of attempts to mitigate bias through technical means (Mehrabi et al., 2022). Unsurprisingly, both argument mining and incivility detection are not immune to the bias problem.

A substantial problem connected to argument mining can arise when the NLP model itself is not equally well trained to recognize texts by different groups, and thus, speech cannot be classified as an argument. For example, NLP models have been shown to be insufficiently trained to handle African American English (AAE) (Blodgett & O'Connor, 2017; Field et al., 2021). Blodgett and O'Connor (2017) show in an analysis of Twitter data that AAE is more often misclassified, for example, as Danish. In the case of NLP, a model is often trained on majority white speech, for example, when Wikipedia is used as a database in training (Field et al., 2021, p. 1907). Models trained on majority white speech can then be ill-equipped to handle such language varieties. AAE, for example, is characterized by several distinct features, such as the use of the habitual be, which indicates reoccurring activity (Green, 2006).[4] Testing the *Stanford part of speech tagger* and *Gate*, Jørgensen et al., (2015, p. 16) found their accuracy to be lacking when applied to AAE and deemed the result "prohibitive of many downstream applications".

However, problems of bias are not limited to the dimension of rationality. In the development of incivility classifiers, the annotation process of manually labelling examples may be crowdsourced, which can introduce the subjective perspective of crowd workers into the annotation process (Sap et al., 2019). This is especially problematic if annotators rate utterances by members of different groups systematically differently, such as when speech containing features of AAE is more often rated as toxic, leading to contributions by African Americans being flagged more often (Sap

[3] The concept of fairness remains the subject of heated debate in the field of AI research (Weinberg 2022). For the purpose of this paper, we use the definition of fairness provided by Mehrabi et al., (2022, p. 2): "In the context of decision-making, fairness is the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics.".

[4] While features such as the habitual be have been observed in many African American communities, it is important not to essentialize features or to disregard regional variety (Wolfram 2007).

et al., 2019). Such discrepancies can, for example, have their roots in AAE markers such as *ass* not denoting a vulgar insult but functioning as a pronoun, such as in *I saw his ass* (Spears, 2021, p. 259). Here it is shown that coders struggle with identifying the context and cultural background necessary to identify incivility (Bormann & Ziegele, 2023, p. 206). Herbst (2010, p. 3) even argues that incivility lies "very much in the eye of the beholder", which presents a problem if the beholder influences how a model operates.

Such problems of biased decisions can lead to unfair outcomes, as participants are excluded based on factors that are not connected to the deliberative quality of their contributions. While AI can potentially promote civility and rationality, AI that functions less well for different groups cannot. However, such problems are not impossible to address. In incivility detection, race and dialect priming can reduce biased labelling (Sap et al., 2019). Furthermore, training data could be supplemented by specifically gathering more data from underrepresented groups in order to improve NLP systems' handling of language varieties (e.g., AAE) (Hovy & Prabhumoye, 2021, p. 6). However, the need to increasingly survey marginalized communities introduces new concerns, such as privacy protection (Nee et al., 2021, p. 5). Generally, efforts are undertaken to debias NLP models and make sure they work equally well for different language varieties and do not display biases in generating tasks (Dacon et al., 2022; Sun et al., 2019). Such interventions primarily treat AI as a technology that can be perfected and debiased, a phrase mentioned in both examples above. For the improvement of online discourses, this would mean that once insufficient training on the varieties of speech and annotator bias are removed, the tools can help to promote deliberative discourse in online spaces. The goal of such interventions is to develop tools that transform input data into output without disadvantaging specific groups. In the following sections we will argue that such a framing focused on the input to output pipeline is insufficient.

5 A Sociotechnical Perspective on AI Tools Promoting Deliberative Norms

Thus far, we have shown that rationality and civility are the most prevalent norms in the context of promoting deliberation through AI tools in online discourse. Additionally, we have analyzed how the complex concepts of those norms are operationalized and, in the process, simplified. Finally, we have discussed how data bias can lead to biased models, influencing the relationship between input data and output. Hence, efforts to cleanse the underlying technology of biases are underway. Drawing on the debate on a sociotechnical approach to AI, we will argue that a focus on the process between input and output ignores the interaction of the technology with inequalities and social hierarchies. We will then apply those insights to the promotion of deliberative norms and show that their promotion through AI tools risks silencing disadvantaged groups by imposing a standard of communication that further benefits privileged groups. This happens because both the norms themselves, but especially their operationalization in AI tools, privilege contributions that follow

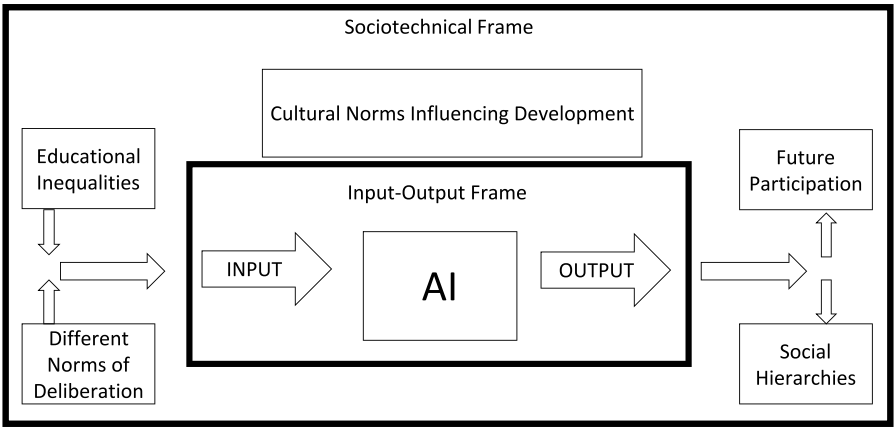

Fig. 1 The sociotechnical and the input–output frame for evaluating AI tools, based on Selbst et al. (2019). *Description:* The figure shows two different boxes, indicating different frames for evaluating AI tools. The inner box, indicates the input–output frame and includes the input data, the AI model and the output data. The outer box encompasses the inner box, and the boxes "cultural norms influencing development", as well as "educational inequalities" and "different norms of deliberation", influencing the input data, which is indicated by arrows towards the inner box. It also includes arrows emerging from the inner box, indicating influence on the boxes "future participation" and "social hierarchies"

a way of expression more highly educated people are better accustomed to while disregarding linguistic, cultural, and socioeconomic variety.

Increasingly, the focus on identifying problems with training data and with the relationship between input and output has come under criticism (e.g., Le Bui & Noble, 2020; Selbst et al., 2019). Occasionally, such criticism has called into question the quest for fair AI in general, for example, when Le Bui and Noble write that "striving for fairness in the face of these systems of power does little to address the ways that digital technologies are increasingly central to other forms of structural power" (2020, p. 178). On the other hand, calls for a re-examination of fairness in the context of AI have been voiced (Barocas & Selbst, 2016).

Selbst et al., (2019, p. 59) argue that attempts to make machine learning algorithms fairer have been limited to considering the relationship between input and output while "abstract[ing] away any context that surrounds this system". This abstraction, shown in Fig. 1, is considered the default without explicitly justifying the choice of this particular framing of the problem (Selbst et al., 2019, p. 60). This means that how the input data comes into being, how human actors engage with the output data, as well as the focus on finding a technical solution at all, remain unquestioned. The process of abstraction fails to consider the interaction of AI with social power structures and leads to the creation of "imagined digital spaces of neutrality and objectivity" (Le Bui & Noble, 2020, p. 171), while ignoring that algorithms are "embedded in social, political, cultural, and economic worlds, shaped by humans" (Crawford, 2021, p. 211). Benjamin (2019) criticizes that ignoring current social power structures in development processes allows those structures to replicate. Brock (2018, p. 1014) additionally criticizes that the development of Information and Communications Technologies often presupposes a norm of a privileged

user, rendering others deficient. Le Bui and Noble (2020, p. 166) criticize attempts to "'unbias' the technology, rather than account for the asymmetrical power relationships and gravity of history". Restricting the analysis of the effects of a tool to input–output relations risks the deepening and reproduction of inequalities by ignoring how the input has been constituted, how the output interacts with the surrounding social system, and which norms shape the perspective of developing AI tools at all. Selbst et al. (2019) thus discuss the concept of a *sociotechnical frame* (see Fig. 1), which explicitly evaluates a model in its interaction with the context it is introduced into.[5] Through such a frame, the inequalities and power structures a model interacts with become a vital part of evaluating a model, and thus, an analysis through this frame may yield different results than an abstract analysis of input–output relations.[6] Consequently, such a frame can contribute to "reshap[ing] the AI ethics research agenda by frontloading the role of power mechanisms" (Gerdes, 2022, p. 4).

If the sociotechnical view is applied to the improvement of online discourse through AI tools, several aspects can be criticized as further excluding marginalized groups that remain beyond the grasp of the input–output frame. We identify three aspects that put the use of AI tools within online discourse at risk of deepening inequalities and exclusion. We will begin with the focus on formal arguments in rationality, then discuss civility and vulgarity, and finally turn to the discouraging effect of the output of a model.

When discussing rationality, it can first be stated that deliberative norms articulate a form of desirable discourse that appears as a neutral baseline to improve discussions, but it is not free from social and cultural norms. As Young (1990, p. 59) puts it, "the dominant group's cultural expressions receive wide dissemination, their cultural expressions become the normal, or the universal". In the face of this supposedly universal standard, other forms of expression are not only rendered as something else but as something inferior or lacking (Young, 1990, p. 116). The ideal of rationality is supposed to lead to the prevailing of the best argument, but in striving for this ideal, other forms of expression are devalued. As Young (1990, p. 118) argues, Habermas focuses on discursive argumentation, excluding other forms of discursive expression, such as metaphors or playful expression, while also maintaining a "dichotomy between reason and affectivity". Such a focus excludes expressions that take the form of humor or narration, which have been discussed under the term type II deliberation (Bächtiger et al., 2010). The operationalization of the Habermasian ideal of rationality only deepens that exclusion, as arguments are supposed to be identified relying on structures of argumentation, automatically ignoring

[5] Note that Selbst et al., (2019, p. 60) also differentiate between the algorithmic frame, exclusively focusing on measures of accuracy, and the data frame, which expands the considerations to, for example, demographic information in the input and output. However, as both framings do not fully capture the social context a model operates in, we have opted here for a binary distinction between an input–output frame and a sociotechnical frame.

[6] For a seminal study of the harm of algorithms developed with insufficient attention to the social context or to the question whether a technical solution is the right choice in the first place see Eubanks (2018).

other contributions such as storytelling, which can be "a way of engaging in argument" (Black, 2008, p. 26). Moreover, this process of formal argumentation is not something that everyone is equally well adapted to. It presupposes familiarity with formal argumentation, privileging those with higher degrees of education, who have had ample practice and familiarity with the exchange of arguments (Sanders, 1997; Young, 2000, p. 38). Sanders (1997, p. 349) even argues that deliberative theory provides seemingly democratic grounds for discrediting the contributions of those who express themselves in different ways, affecting especially the systematically disadvantaged. Habermas himself conceptualizes the deliberative norms as an ideal, but having acquired education clearly makes it easier to come close to this ideal. When such an ideal is taken to be a neutral standard, groups that have privileged access to power, resources, and education that makes them uniquely suited to follow those norms can more effectively pursue their interests and seemingly neutral procedures "will […] yield outcomes in the interests of the more powerful" (Young, 1990, p. 114). In societies in which access to education correlates with race, generational poverty, and affluence, such an ideal of rational argument can work to exclude, as some are "more learned and practiced in making arguments" (Sanders, 1997, p. 349). Consequently, Young (2000, pp. 39–40) sees in the ideal of a rational dispassionate communication the "speech culture of white middle-class men" while the "speech culture of women, racialized or ethicized minorities, and working-class people, on the other hand, often is, or is perceived to be, more excited and embodied, values more the expression of emotion". What appears as a neutral ideal is thus revealed to be deeply influenced by value judgements of what communication should look like. Those judgements can disproportionately burden already disadvantaged groups, especially through privileging specific argumentative structures in the operationalization of rationality. Implemented as an AI model in online discourse, the enforcement of this standard of rationality may be perceived as neutral moderation for the sake of a better discussion, leaving aside the ways cultural norms, value judgements and power structures have influenced the pipeline from conceptualization to actual programming decisions.

Second, this problem of norm setting also becomes apparent when looking at the norm of civility and, most poignantly, to vulgarity as a marker of incivility. We have already discussed that the annotation process for vulgarity risks coding elements of AAE as vulgar (Sap et al., 2019; Spears, 2021). Here, the line between a flawed annotation process and dominant cultural norms working to exclude people blurs, as the speech in question is understood as non-toxic and not uncivil by AAE speakers (Spears, 2021). But even if members of some social groups used vulgar expression more often than others, this can be understood more as a difference than a deficiency. In his classic study of black urban youth vernacular, Labov (1972) shows that insults serve a ritualistic bonding function within some peergroups and are used to demonstrate high language skill. Civility as a norm carries echoes of what Young (1990, p. 110) calls the old ideal of respectability, which describes the virtuous man as "unyielding to passion". Expressions are then governed by this standard as "some words are clean and respectable, others are dirty" (Young, 1990, p. 137). If such an ideal shows itself in the filtering of vulgarity, it presupposes a norm of a restrained discourse that

again leaves little room for different behavior. Moreover, it ignores that a standard of acceptable language cannot be separated from the cultural and socioeconomic factors that have shaped it. Historically, the charge of incivility has often been leveled against disenfranchised minorities (Zerilli, 2014, p. 108). While hate speech and attempts to silence are rightly condemned, vulgarity filters come closer to an "identification of reasonable open public debate with polite, orderly, dispassionate, gentlemanly argument" (Young, 2000, p. 49). This is especially problematic since disruptive communication can provide an effective way to call attention to issues affecting groups that hold little power and to make their views part of the political discourse (Young, 2000, p. 50). Thus, civility is not a neutral norm, and its operationalization in AI tools risks filtering out voices that have historically been silenced. It thus risks automating the enforcement of "the communicative styles of already powerful groups" (Bickford, 2011, p. 1025). Once again, this problem is especially pronounced because the operationalization of the deliberative norm is simplified for the deployment of AI tools. Complex arguments about the role of respect are foregone, and simple name-calling and vulgarity are filtered, even though the use of such elements may vary widely between different groups, may fulfill different linguistic functions and depends on the socioeconomic class of the participants.

Third, the norms that are enforced through AI tools do not only devalue some contributions but can also be expected to shape the behavior of potential participants when considering future contributions. When moving beyond the relationship between input and output, it should also be questioned how the output could shape future input. This means asking, how highlighting a specific kind of argument or privileging a certain kind of contribution in online discussions will shape future contributions to discourse. If the highlighted arguments in an online discussion all fulfil criteria of formal arguments, will those whose contribution would have looked different still decide to contribute? As those systems become more widespread, it is possible that people self-censor, in order to avoid vulgarity filters and other tools, thus curtailing the variety of expressions.

Those problems with deliberative norms and their operationalization only come into view when the scope of analyzing AI tools is broadened beyond the relationship between input and output. Questioning the cultural and social norms as well as the social hierarchies that underlie the ideal of a good discourse means venturing beyond scrutinizing whether criteria for a good discourse are fairly applied. Instead, it means scrutinizing how the input comes into being in the first place and how educational, cultural, and socioeconomic differences shape the style of argumentation or the use of vulgarity. While attempting to improve online discourses, AI tools can potentially deepen inequalities in access to political participation and discussion and exclude disadvantaged groups from another aspect of public life. On the other side, those with more access to educational resources will be able to pursue their interests more effectively in a seemingly neutral process. However, uncovering those problems requires leaving behind the idea that AI tools can be neutral (Le Bui & Noble, 2020, p. 171).

6 The Choice of the Frame and Objectionable Indifference

In what way interactions between an AI tool and the surrounding social context and hierarchies are considered ultimately constitutes a choice between different framings of the problem. In this final part, we present a normative argument for a sociotechnical frame as the preferable option when it comes to AI tools for online discourse. We argue that choosing a narrow frame presents a form of objectionable indifference, as it expresses that the predictable disadvantages incurred by some groups are not a matter of concern. After rejecting the challenge that a sociotechnical frame unduly leads to a more complex development process, we argue that tools for online discussions should be explicitly scrutinized for their potential to interact with social hierarchies. Such scrutiny is likely to influence how and even if certain AI tools are developed and deployed. Finally, we critically discuss the fact that AI tools for equality and reciprocity receive comparatively little attention.

We borrow the concept of objectionable indifference from Sangiovanni (2017), who develops it in his theory of discrimination. He argues that actions can inferiorize and demean those affected, depending on the attitude of those carrying out an action and the social meaning attached to an action (Sangiovanni, 2017). Social meaning can only be analyzed by looking at the message an action or policy sends "against a wider social, cultural, political, economic background" (Sangiovanni, 2017, pp. 122–123). Hence, an action's potential to inferiorize and demean depends on the possibility to mobilize or deepen patterns of stigma and exclusion (Sangiovanni, 2017, p. 139), making members of disadvantaged groups uniquely vulnerable, a stance that Sangiovanni shares with other theorists of discrimination (e.g., Hellman, 2011). Finally, inaction in the face of the reproduction of patterns of stigma and inequality expresses that those patterns are acceptable or, at the very least, not a matter of weighty concern, thus expressing objectionable indifference (Sangiovanni, 2017, pp. 168–171). This framework renders the continuation of patterns of exclusion a choice instead of a regrettable inevitability and allows for the complexity of social hierarchies and cultural norms to play a crucial role in the analysis of an action. When patterns of exclusion are reproduced, it is expressed that such patterns are acceptable. This concept ascribes an active role to actors who deepen or reproduce social inequalities and makes their actions open to analysis in light of their alternative choices. We will apply this conceptual framework to the framing of AI models in online discussions.

The exclusion of alternative forms of argumentation, enforcement of a privileged perspective on civility, and subsequently the further exclusion of marginalized groups from another aspect of social participation are not inevitable. Rather, it is a deliberate design choice that only appears as inevitable through a narrow focus on the input–output relationship. As a choice, the setting of a specific frame can be analyzed for its social meaning. For the first element of this analysis, it is important to state that negative effects on disadvantaged groups are to be expected if a narrow input–output frame is chosen. As Bickford (2011, p. 1025) phrases it concerning the operationalization of rationality, "inegalitarian effects

are inevitable unless we expand our understanding of political communication to include more affective elements". As we have seen, rationality, operationalized as the use of structured argument, and civility, marking out vulgarity as uncivil, can be expected to benefit the privileged and enforce specific cultural norms. Furthermore, this ties in with a history of silencing, ignoring, and policing the expressions of disadvantaged groups (Zerilli, 2014, p. 108). Considering AI tools in online discourses as potentially neutral against this background ignores the multiple ways such technologies interact with inequalities and social hierarchies. As a result, patterns of exclusion and stigmatization, as well as disregard for language varieties and forms of expression deviating from the norms of privileged groups, are likely to be reproduced. By choosing a narrow frame of investigation such effects are, by definition, no longer a concern. Choosing such a narrow frame thus communicates indifference, as the interactions between AI tools and social structures are simply considered irrelevant to the design of such tools, even though they are likely to occur if they are not explicitly addressed. Moreover, by excluding other forms of argumentation and by adapting a culturally laden standard of civility, historical patterns of exclusion and stigmatization are continued and disadvantaged groups experience further exclusion from participation. In front of this historical and cultural background, the choice to limit concerns to an input–output relationship communicates that those effects are acceptable, and thus, this choice communicates objectionable indifference. Enlarging the frame, however, would allow developers to recognize that AI tools should be developed with an understanding of the diversity of real discourse. For example, aspects of type II deliberation, such as narratives, emotions, or testimony (Bächtiger et al., 2010), as well as different language varieties and sociolects, should be explicitly considered in the development process.

Such an argument can be expected to run into the objection that it unduly expands the focus of AI development. While the effects described above may be regrettable, AI developers and researchers should not be made responsible for what Le Bui and Noble (2020, p. 166) call the "gravity of history". They have not caused the educational inequalities that render some people more adapted to a formalized style of argumentation. Taking into account the interactions between inequalities that exist in society and AI tools would simply make the process of deploying any tool so complex that such tools would be much harder to develop and deploy. This challenge to the argument, however, does not succeed. First, it is widely accepted within the community of researchers and developers of AI tools that increased complexity of the development process has to be accepted to avoid harmful outcomes in some instances. Mitigating the effect of bias that is introduced through the training data requires additional testing and a close analysis of the training data (Leavy et al., 2020). Here, the trade-off between mitigating negative effects on disadvantaged groups and engaging in a more complex research and development process is already accepted. It remains true that the inequalities reproduced or deepened through AI tools are not caused by the deployment of said tools. Nonetheless, the technology interacts with the social structures it is introduced into. Any effects that occur due to this interaction result partially from introducing the technology into an existing social context where it will predictably reproduce patterns of inequality

and uphold dominant cultural norms. To analyze such effects beforehand, impact assessment for AI that takes into account multiple stakeholders as well as the specific organizational and social context of a specific use case can provide a pre-emptive safeguard against harmful effects (Stahl et al., 2023, p. 1281). In the case of AI online discussions, this would include identifying the groups most at risk of being excluded as well as the impact that different definitions and operationalizations of deliberative norms are likely to have on them. Furthermore, the existing inequalities render it necessary to pay special attention to disadvantaged groups, so that disadvantages incurred by them can not simply be compensated by advantages for more privileged groups. The philosophical concept of prioritarianism, in which the interests of the least well-off are paid special attention, can provide a starting point for such assessment (see Parfit, 2012). Against this background, limiting the frame to an input–output relationship constitutes an arbitrary choice that expresses indifference to the wider effects that the introduction of a tool will predictably cause. This analysis, using objectionable indifference as a framework, provides an explicit normative justification for adopting a sociotechnical frame, complementing works that point out the effects of different frames (see for example Selbst et al., 2019).

Considering problems that stem from the interactions of algorithms and social systems will lead to some AI tools not being developed or deployed. We argue that this is actually desirable. Selbst et al., (2019, p. 63) describe the idea that a technical solution to a problem must exist as the solutionism trap and argue that "to understand whether to build, we must also understand the existing social system". If the goal is to make online publics a better place for democratic discourse, this goal is simply not satisfied if the resulting tools end up increasing the advantages of the already privileged. In some cases, it is necessary to resist the "flattening of complexity into clean signal for the purposes of prediction" (Crawford, 2021, p. 213). This means recognizing that sometimes the technological tools available are simply insufficient to deal with the complexity of real discourse, instead of falling into the solutionism trap. Pressing ahead with AI tools despite the problems that come into view by adopting a sociotechnical view expresses that the costs incurred by disadvantaged groups are acceptable. Either those problems must be resolved through changes to the operationalization deliberative norms or certain tools ought not to be deployed.

This, however, does not mean that promoting deliberative norms in online discourses should be abandoned or that narration or different standards of civility can never be excluded without expressing objectionable indifference. The setting of an online discourse may influence how deliberative norms potentially work to exclude some participants. Consider the example of a faculty holding an online discussion about new PhD guidelines (Friess, 2018). Here, the participants can likely be expected to be familiar with structured argument in the way the deliberative norm presupposes. Hence, the criticism presented does not apply to this specific situation. Such specifications would mean adapting AI tools to specific contexts, instead of developing code that is as portable as possible (Selbst et al., 2019). Furthermore, to identify the needs and competencies of potential participants in a specific online discussion, it is necessary to involve a more diverse range of people, beyond programmers and other experts, throughout the development in a process of "co-creation" (Sartori & Theodorou, 2022, p. 8). In online discussions that are supposed to further

democratic discourse, diversity, both in the voices influencing development as well as within the discourse, can be seen as a value in itself, rendering the discourse more representative. The field of AI research for online discussions could take inspiration from the research on inclusive news recommenders, where approaches to measure and increase the visibility of minority voices have already been made (see Helberger, 2019, p. 1006; Vrijenhoek et al., 2021, pp. 178–180). This approach means taking seriously all effects a model introduced into a specific social context may cause. Which kind of argumentation to include and which understanding of rationality or civility to adopt are fundamental choices that can influence which voices are given more visibility and should be addressed and evaluated as such through a sociotechnical frame. Otherwise, some negative effects are simply rendered irrelevant, expressing objectionable indifference towards those bearing the brunt of those effects.

Ultimately, this also brings into critical view the state of AI online deliberation research itself. Equality and reciprocity, which explicitly aim at promoting more engagement by a more diverse group of participants, are insufficiently addressed by AI research compared to rationality and civility. A development and research process focused on co-creation and representation of marginalized voices (Sartori & Theodorou, 2022) may lead to greater engagement with those norms. However, such a development process is complicated and requires engaging different groups of users and ultimately depends on their cooperation and participation. Structuring the existing discourse through enforcing relatively formal norms of rationality and civility lends itself better to being addressed through a technological solution. However, if feasibility becomes the driver of the direction that AI for online discussions research takes, there is a danger of letting the development of technology also set the research agenda on how to foster better and more inclusive discourses. As Crawford (2021, p. 214) argues, the development of AI should not "focus on the innovative nature of the method rather than on what is primary: the purpose of the thing itself". A sociotechnical view can help to critically evaluate such larger trends beyond specific use cases and applications. The approach can shine a light on the possibility that marginalized groups may remain underconsidered in the development of AI tools for online discussions generally, not only in operationalizing rationality and civility. It should be assessed how a representative discourse that engages a wide variety of participants can be facilitated, and it should be openly analyzed which and if AI tools can contribute to reaching this goal instead of focusing on norms that can be addressed relatively easily through existing technology.

7 Conclusion

This paper aims to provide ethical reflections and a critical sociotechnical perspective on AI interference in online discussions. Discussing the state of online discourse, we have shown that norms of deliberation are still considered to provide an important normative standard to evaluate such discourse. This also translates into AI research mainly focusing on rationality and civility, both key norms of deliberation. We have shown that those norms are often operationalized as identification

of argumentative structures for rationality and vulgarity detection for civility in the context of AI tools. We have analyzed, through a sociotechnical frame, how such AI tools for online discourse can contribute to patterns of exclusion and inequality. Moreover, we have adopted the concept of objectionable indifference to provide a normative argument for a sociotechnical framing in the discussion of AI tools in online discussions. Finally, we have suggested possible approaches to developing AI tools following a sociotechnical frame. Such approaches should explicitly assess the impact of an AI tool on marginalized and disadvantaged groups, considering the context the tool will operate in. Furthermore, such groups should be included in the development process, and the communication styles of different groups should be recognized by the final product. A promising step in developing a context-sensitive AI tool for civility can be found in the work of Arora et al. (2020), who propose to work with communities most strongly affected by online abuse in the annotation process when developing harassment and hate speech classifiers. Such an approach takes into account the targeting of specific groups through online harassment, which itself is often based on power differentials within the online community. Furthermore, it could capture the nuances of abuse targeted at marginalized groups instead of broadly relying on crowd workers to develop a tool promoting civility regardless of context. Generally, inclusivity and representation ought to be taken into account more strongly as goals of tools for the improvement of online discussions. Finally, if a sociotechnical frame reveals AI tools to be increasing inequalities, ceasing the development of such tools should be considered as a real alternative to techno-centric solutionism.

Against the backdrop of a sociotechnical perspective, the contribution of this paper is threefold: First, we have shown that AI tools for online discourse mostly focus on relatively simple operationalizations of the complex concepts of rationality and civility. Second, we have shown in the context of online deliberation that adopting a sociotechnical framing yields a deeper analysis of the ways simple operationalizations of rationality and incivility deepen and reproduce inequalities. Third, we have gone beyond stating that a sociotechnical framing delivers different results than an input–output frame and provided a normative argument showing that adopting a sociotechnical frame in the context of online deliberation is ethically preferable.

However, there are several limitations to this paper. While we have provided some indications on how a sociotechnical frame could serve as a basis for the development of more inclusive AI tools, we have only provided limited suggestions on concrete technical steps. Further research may fill this gap by considering our ethical arguments when designing AI tools in the context of online deliberation. Furthermore, in this article, we adopted a deliberative perspective in order to critique the operationalization of certain deliberative norms. However, this criticism of deliberative norms is not new (see: Sanders, 1997; Young, 2000), but simply adapted to the context of AI tools in online discussions. Nevertheless, other theoretical perspectives are also promising. For example, agonistic approaches (e.g., Mouffe, 2005) could offer a fruitful alternative to our approach based on deliberative theory since they focus on conflict and fundamental differences between groups. In this context, sociotechnical approaches could lay bare fundamental differences of interests between groups (Sartori & Theodorou, 2022), instead of suggesting that those interests could be easily

reconcilable if only a technical solution for online discourse was developed. Combining critical sociotechnical approaches with agonistic approaches to communication and AI development could help to "call out the ethical and political implications of *who decides task T*, *performance metric P* and *experience E*, and to investigate *how this is done*, taking into account *which (and whose) concerns are at stake*" (Hildebrandt, 2019, p. 110, cursive in the original). Finally, the paper has focused on rationality and civility while providing only limited engagement with equality and reciprocity. We justified our focus with regard to the current state of AI research, which has mainly focused on rationality and civility (Friess et al., 2022). However, this selectivity of previous research is not a strong argument for omitting already neglected aspects. Nevertheless, our paper is intended to highlight the narrowing of AI deliberation research to rationality and civility as seemingly easy to operationalize concepts. Those two dimensions and their use in AI development risk the exclusion of minority groups because they formalize a particular style of communication and a particular kind of argument or contribution. Those norms are therefore both dominant in the research field and particularly problematic, which is why a critical reflection of their impact is particularly important. Nonetheless, this paper suggests that more research on AI tools promoting the dimensions of equality and reciprocity is desirable. Such work should also scrutinize the pipeline from conceptualization to operationalization, to the use in AI tools for those norms.

Acknowledgements We would like to express our gratitude to the two anonymous reviewers. Their constructive criticism and suggestions were very helpful and improved our manuscript.

Author Contributions Dr. Dennis Friess has contributed the introduction, as well as section one and two. He is also one of the authors who conducted the systematic literature review. Moreover, the original idea to collaborate has followed from a talk Dr. Dennis Friess has given. Jonas Aaron Carstens has contributed section three, four and five, as well as the conclusion. Both authors reviewed and provided feedback for each other's contributions.

Funding Open Access funding enabled and organized by Projekt DEAL. This research has benefited from generous funding by the Jürgen Manchot Foundation through the Manchot research group "Decision-making with the help of Artificial Intelligence".

Declarations

Conflict of interest The authors declare that there is no conflict of interest or financial ties to disclose.

References

Aitamurto, T., & Landemore, H. (2013). Democratic Participation and Deliberation in Crowdsourced Legislative Processes: The Case of the Law on Off-Road Traffic in Finland. In *The 6th Conference on Communities and Technologies (C&T), Workshop: Large-Scale Idea Management and Deliberation Systems.*

Anderson, A. A., Brossard, D., Scheufele, D. A., Xenos, M. A., & Ladwig, P. (2014). The "Nasty Effect:" Online incivility and risk perceptions of emerging technologies. *Journal of Computer-Mediated Communication, 19*(3), 373–387. https://doi.org/10.1111/jcc4.12009

Anderson, A. A., Yeo, S. K., Brossard, D., Scheufele, D. A., & Xenos, M. A. (2018). Toxic Talk: How online incivility can undermine perceptions of media. *International Journal of Public Opinion Research, 30*(1), 156–168. https://doi.org/10.1093/ijpor/edw022

Argyle, L. P., Bail, C. A., Busby, E. C., Gubler, J. R., Howe, T., Rytting, C., Sorensen, T., & Wingate, D. (2023). Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.2311627120

Arora, I., Guo, J., Levitan, S. I., McGregor, S., & Hirschberg, J. (2020). A novel methodology for developing automatic harassment classifiers for Twitter. In S. Akiwowo, B. Vidgen, V. Prabhakaran, & Z. Waseem (Eds.), *Proceedings of the fourth workshop on online abuse and harms* (pp. 7–15). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.alw-1.2

Bächtiger, A., Niemeyer, S., Neblo, M., Steenbergen, M., & Steiner, J. (2010). Disentangling diversity in deliberative democracy: Competing theories, their blind sport and complementarities. *Journal of Political Philosophy, 18*, 32–63. https://doi.org/10.1111/j.1467-9760.2009.00342.x

Bächtiger, A., & Parkinson, J. (2019). *Mapping and measuring deliberation: Towards a new deliberative quality*. Oxford University Press.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*(3), 671–732. https://doi.org/10.15779/Z38BG31

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim Code*. Polity Press.

Bickford, S. (2011). Emotion talk and political judgment. *The Journal of Politics, 73*(4), 1025–1037. https://doi.org/10.1017/S0022381611000740

Black, L. W. (2008). Listening to the city: Difference, identity, and storytelling in online deliberative groups. *Journal of Deliberative Democracy*. https://doi.org/10.16997/jdd.76

Blodgett, S. L., & O'Connor, B. (2017). Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English. In *2017 Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2017)*. https://arxiv.org/pdf/1707.00061

Bormann, M., & Ziegele, M. (2023). Incivility. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 199–217). https://doi.org/10.48541/dcr.v12.12

Brock, A. (2018). Critical technocultural discourse analysis. *New Media & Society, 20*(3), 1012–1030. https://doi.org/10.1177/1461444816677532

Brüggemann, M., Elgesem, D., Bienzeisler, N., Dedecek-Gertz, H., & Walter, S. (2020). Mutual group polarization in the blogosphere: Tracking the hoax discourse on climate change. *International Journal of Communication, 14*, 1025–1048.

Buchstein, H. (1996). Bittere Bytes [Bitter Bytes]. *Deutsche Zeitschrift Für Philosophie, 44*(4), 583–607.

Cabrio, E., & Villata, S. (2018). Five years of argument mining: A Data-driven analysis. *IJCAI, 18*, 5427–5433. https://doi.org/10.24963/ijcai.2018/766

Chadwick, A. (2009). Web 2.0: New challenges for the study of E-democracy in an era of informational exuberance. *I/S: A Journal of Law and Policy for the Information Society, 5*(1), 9–41.

Clever, L., Klapproth, J., & Frischlich, L. (2022). Automatisierte (Gegen-)Rede? Social Bots als digitales Sprachrohr ihrer Nutzer*innen [Automated (Counter-)Speech? Social bots as a digital mouthpiece of their users]. In J. Ernst, M. Trompeta, & H.-J. Roth (Eds.), *Gegenrede digital: Neue und alte Herausforderungen interkultureller Bildungsarbeit in Zeiten der Digitalisierung* (pp. 11–26). Springer.

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication, 64*(4), 658–679. https://doi.org/10.1111/jcom.12104

Coleman, S., & Gøtze, J. (2001). *Bowling together. Online public engagement in policy deliberation*. Hansard Society.

Crawford, K. (2021). *Atlas of AI*. Yale University Press. https://doi.org/10.12987/9780300252392

Curiskis, S. A., Drake, B., Osborn, T. R., & Kennedy, P. J. (2020). An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. *Information Processing & Management*. https://doi.org/10.1016/j.ipm.2019.04.002

Dacon, J., Liu, H., & Tang, J. (2022). Evaluating and Mitigating Inherent Linguistic Bias of African American English through Inference. In *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 1442-1454). International Committee on Computational Linguistics

Dahlberg, L. (2001). Extending the public sphere through cyberspace: The case of Minnesota E-Democracy. *First Monday, 6*(3), 1–8. https://doi.org/10.5210/fm.v6i3.838

Dahlberg, L. (2007). The Internet, deliberative democracy, and power: Radicalizing the public sphere. *International Journal of Media & Cultural Politics, 3*(1), 47–64. https://doi.org/10.1386/macp.3.1.47_1

Dahlberg, L. (2010). The internet and democratic discourse: Exploring the prospects of online deliberative forums extending the public sphere. *Information, Communication & Society, 4*(4), 615–633. https://doi.org/10.1080/13691180110097030

Davidson, S., Sun, Q., & Wojcieszak, M. (2020). Developing a New Classifier for Automated Identification of Incivility in Social Media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms.* Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.alw-1.12

Davis, M. (2021). The online anti-public sphere. *European Journal of Cultural Studies, 24*(1), 143–159. https://doi.org/10.1177/1367549420902799

Daxenberger, J., Ziegele, M., Gurevych, I., & Quiring, O. (2018). Automatically Detecting Incivility in Online Discussions of News Media. In *2018 IEEE 14th International Conference on e-Science* (pp. 318–319). IEEE. https://doi.org/10.1109/eScience.2018.00072

Dutta, S., Juneja, J., Das, D., & Chakraborty, T. (2022, March 24). *Can Unsupervised Knowledge Transfer from Social Discussions Help Argument Mining?* arXiv preprint. https://arxiv.org/pdf/2203.12881

Dutwin, D. (2003). The character of deliberation: Equality, argument, and the formation of public opinion. *International Journal of Public Opinion Research, 15*(3), 239–264. https://doi.org/10.1093/ijpor/15.3.239

Esau, K., Friess, D., & Eilders, C. (2017). Design Matters! An empirical analysis of online deliberation on different news platforms. *Policy & Internet, 9*(3), 321–342. https://doi.org/10.1002/poi3.154

Eubanks, V. (2018). *Automating inequality: How high tools profile, police, and punish the poor* (First Picador). St. Martin's Press.

Field, A., Blodgett, S. L., Waseem, Z., & Tsvetkov, Y. (2021). A Survey of Race, Racism, and Anti-Racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics.* http://arxiv.org/pdf/2106.11410v2

Friess, D., Weinmann, C., & Behrendt, M. (2022). AI and Deliberation. How AI can Support Online Discussions in Deliberative Fashion – a Systematic Review. In *9th European Communication Conference, Aarhus.* https://diid.hhu.de/wp-content/uploads/2024/01/Friess-et-al.-2022_ECREA_AI-Deliberation-Review.pdf

Friess, D. (2018). Letting the faculty deliberate: Analyzing online deliberation in academia using a comprehensive approach. *Journal of Information Technology & Politics, 15*(2), 155–177. https://doi.org/10.1080/19331681.2018.1460286

Friess, D., & Eilders, C. (2015). A systematic review of online deliberation research. *Policy & Internet, 7*(3), 319–339. https://doi.org/10.1002/poi3.95

Friess, D., Ziegele, M., & Heinbach, D. (2021). Collective civic moderation for deliberation? Exploring the links between citizens' organized engagement in comment sections and the deliberative quality of online discussions. *Political Communication, 38*(5), 624–646. https://doi.org/10.1080/10584609.2020.1830322

Fu, Y., Wang, S., Li, X., Li, D., Li, Y., Liao, J., & Zheng, J. (2023). Hierarchical neural network: Integrate divide-and-conquer and unified approach for argument unit recognition and classification. *Information Sciences, 624*, 796–810. https://doi.org/10.1016/j.ins.2022.12.050

Garland, J., Ghazi-Zahedi, K., Young, J.-G., Hébert-Dufresne, L., & Galesic, M. (2022). Impact and dynamics of hate and counter speech online. *EPJ Data Science*. https://doi.org/10.1140/epjds/s13688-021-00314-6

Gerdes, A. (2022). The tech industry hijacking of the AI ethics research agenda and why we should reclaim it. *Discover Artificial Intelligence*. https://doi.org/10.1007/s44163-022-00043-3

Graham, T., & Witschge, T. (2003). In search of online deliberation: Towards a new method for examining the quality of online discussions. *Communications, 28*(2), 173–204. https://doi.org/10.1515/comm.2003.012

Green, L. (2006). African American English. In E. Finegan & J. R. Rickford (Eds.), *Language in the USA: Themes for the twenty-first century* (pp. 76–91). Cambridge University Press.

Gutmann, A., & Thompson, D. F. (2004). *Why deliberative democracy?* Princeton University Press.

Habermas, J. (1983). *Moralbewußtsein und kommunikatives Handeln [Moral Consciousness and Communicative Action]*. Suhrkamp.

Habermas, J. (1995). *Theorie des Kommunikativen Handelns [The Theory of Communicative Action]*. Suhrkamp.

Habermas, J. (1998). *Faktizität und Geltung: Beiträge zur Diskurstheorie des Rechts und des demokratischen Rechtsstaats [Between facts and norms. Contributions to a discourse theory of law and democracy]*. Suhrkamp.

Habernal, I., & Gurevych, I. (2017). Argumentation mining in user-generated web discourse. *Computational Linguistics, 43*(1), 125–179. https://doi.org/10.1162/COLI_a_00276

Hancock, J. T., Naaman, M., & Levy, K. (2020). AI-Mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication, 25*(1), 89–100. https://doi.org/10.1093/jcmc/zmz022

Helberger, N. (2019). On the democratic role of news recommenders. *Digital Journalism, 7*(8), 993–1012. https://doi.org/10.1080/21670811.2019.1623700

Hellman, D. (2011). *When is discrimination wrong?* Harvard University Press.

Herbst, S. (2010). *Rude democracy: Civility and incivility in American politics*. Temple University Press.

Hildebrandt, M. (2019). Privacy as protection of the incomputable self: From agnostic to agonistic machine learning. *Theoretical Inquiries in Law, 20*(1), 83–121. https://doi.org/10.1515/til-2019-0004

Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass, 15*(8), 1–19. https://doi.org/10.1111/lnc3.12432

Hsueh, M., Yogeeswaran, K., & Malinen, S. (2015). "Leave your comment below": Can biased online comments influence our own prejudicial attitudes and behaviors? *Human Communication Research, 41*(4), 557–576. https://doi.org/10.1111/hcre.12059

Hwang, H., Kim, Y., & Huh, C. U. (2014). Seeing is believing: Effects of uncivil online debate on political polarization and expectations of deliberation. *Journal of Broadcasting & Electronic Media, 58*(4), 621–633. https://doi.org/10.1080/08838151.2014.966365

Ida, M., Morio, G., Iwasa, K., Tatsumi, T., Yasui, T., & Fujita, K. (2019). Can You Give Me a Reason? Argument-inducing Online Forum by Argument Mining. In L. Liu (Ed.), *ACM Digital Library, The World Wide Web Conference* (pp. 3545–3549). Association for Computing Machinery. https://doi.org/10.1145/3308558.3314127

Ito, T., Hadfi, R., & Suzuki, S. (2022). An agent that facilitates crowd discussion. *Group Decision and Negotiation, 31*(3), 621–647. https://doi.org/10.1007/s10726-021-09765-8

Jiang, H., & Nachum, O. (2020). Identifying and Correcting Label Bias in Machine Learning. *International Conference on Artificial Intelligence and Statistics*, 702–712. http://proceedings.mlr.press/v108/jiang20a.html

Jo, Y., Bang, S., Manzoor, E., Hovy, E., & Reed, C. (2020). Detecting Attackable Sentences in Arguments. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1–23). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.1

Jørgensen, A., Hovy, D., & Søgaard, A. (2015). Challenges of studying and processing dialects in social media. In *Proceedings of the Workshop on Noisy User-generated Text* (pp. 9–18). Association for Computational Linguistics. https://doi.org/10.18653/v1/w15-4302

Kim, S., Eun, J., Seering, J., & Lee, J. (2021). Moderator chatbot for deliberative discussion. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), 1–26. https://doi.org/10.1145/3449161

Klein, M. (2022). *Crowd-scale deliberation for complex problems: A progress report*. Advance online publication. https://doi.org/10.2139/ssrn.4049680

Kreissel, P., Ebner, J., Urban, A., & Guhl, J. (2018). *Hass auf Knopfdruck Rechtsextreme Trollfabriken und das Ökosystem koordinierter Hasskampagnen im Netz [Hate at the press of a button, trollfarms and the ecosystem of coordinated hate-campaigns on the internet]*. Institute for Strategic Dialogue.

Labov, W. (1972). *Language in the inner city: Studies in the Black English vernacular*. University of Pennsylvania Press.

Lawrence, J., Park, J., Budzynska, K., Cardie, C., Konat, B., & Reed, C. (2017). Using argumentative structure to interpret debates in online deliberative democracy and eRulemaking. *ACM Transactions on Internet Technology, 17*(3), 1–22. https://doi.org/10.1145/3032989

Lawrence, J., & Reed, C. (2020). Argument mining: A survey. *Computational Linguistics, 45*(4), 765–818. https://doi.org/10.1162/coli_a_00364

Le Bui, M., & Noble, S. U. (2020). We're missing a moral framework of justice in artificial intelligence. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *Oxford Handbooks. The Oxford handbook of ethics of AI* (pp. 161–179). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.9

Leavy, S., Meaney, G., Wade, K., & Greene, D. (2020). Mitigating gender bias in machine learning data sets. In L. Boratto, S. Faralli, M. Marras, & G. Stilo (Eds.) *Communications in computer and information science* (pp. 12–26). Springer International Publishing. https://doi.org/10.1007/978-3-030-52485-2_2

Liebeck, M., Esau, K., & Conrad, S. (2016). What to Do with an Airport? Mining Arguments in the German Online Participation Project Tempelhofer Feld. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016).* Association for Computational Linguistics. https://doi.org/10.18653/v1/w16-2817

Lippi, M., & Torroni, P. (2015). Argument mining: A machine learning perspective. In E. Black, S. Modgil, & N. Oren (Eds.), *Lecture notes in computer science. Theory and applications of formal argumentation* (pp. 163–176). Springer International Publishing. https://doi.org/10.1007/978-3-319-28460-6_10

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys, 55*(9), 1–35. https://doi.org/10.1145/3560815

Lowry, S., & Macpherson, G. (1988). A blot on the profession. *British Medical Journal (clinical Research Ed.), 296*(6623), 657–658. https://doi.org/10.1136/bmj.296.6623.657

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), 1–35. https://doi.org/10.1145/3457607

Moreau, S. (2020). *Faces of Inequality: A Theory of Wrongful Discrimination*. Oxford University Press.

Mouffe, C. (2005). *On the political (Thinking in action)*. Routledge.

Nee, J., Macfarlane Smith, G., Sheares, A., & Rustagi, I. (2021). Advancing social justice through linguistic justice: Strategies for building equity fluent NLP technology. *Equity and access in algorithms, mechanisms, and optimization* (pp. 1–9). ACM. https://doi.org/10.1145/3465416.3483301

Negroponte, N. (1995). *Being digital*. Alfred A. Knopf

Nelimarkka, M., Nonnecke, B., Krishnan, S., Aitamurto, T., Catterson, D., Crittenden, C., Garland, C., Gregory, C., Huang, Ching-Chang, Newsom, G., Patel, J., Scott, J., & Goldberg, K. (2014). *Comparing Three Online Civic Engagement Platforms using the Spectrum of Public Participation.* UC Berkeley: Center for Information Technology Research in the Interest of Society (CITRIS).

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., & Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. https://doi.org/10.1002/widm.1356

Pacheco, D., Flammini, A., & Menczer, F. (2020). Unveiling Coordinated Groups Behind White Helmets Disinformation. In A. E. F. Seghrouchni, G. Sukthankar, T.-Y. Liu, & M. van Steen (Eds.), *Companion Proceedings of the Web Conference 2020* (pp. 611–616). ACM. https://doi.org/10.1145/3366424.3385775

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society, 6*(2), 259–283. https://doi.org/10.1177/1461444804041444

Parfit, D. (2012). Another defence of the priority view. *Utilitas, 24*(3), 399–440. https://doi.org/10.1017/S095382081200009X

Rodríguez-Ruiz, J., Mata-Sánchez, J. I., Monroy, R., Loyola-González, O., & López-Cuevas, A. (2020). A one-class classification approach for bot detection on Twitter. *Computers & Security*. https://doi.org/10.1016/j.cose.2020.101715

Romberg, J., & Conrad, S. (2021). Citizen Involvement in Urban Planning - How Can Municipalities Be Supported in Evaluating Public Participation Processes for Mobility Transitions? In K. Al-Khatib, Y. Hou, & M. Stede (Eds.), *Proceedings of the 8th Workshop on Argument Mining* (pp. 89–99). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.argmining-1.9

Ruckdeschel, M., & Wiedemann, G. (2022). Boundary Detection and Categorization of Argument Aspects via Supervised Learning. In *Proceedings of the 9th Workshop on Argument Mining* (pp. 126–136).

Ruiz, C., Domingo, D., Micó, J. L., Díaz-Noci, J., Meso, K., & Masip, P. (2011). Public Sphere 2.0? The democratic qualities of citizen debates in online newspapers. *The International Journal of Press/politics, 16*(4), 463–487. https://doi.org/10.1177/1940161211415849

Sadeque, F., Rains, S., Shmargad, Y., Kenski, K., Coe, K., & Bethard, S. (2019). Incivility Detection in Online Comments. In R. Mihalcea, E. Shutova, L.-W. Ku, K. Evang, & S. Poria (Eds.), *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)* (pp. 283–291). Association for Computational Linguistics. https://doi.org/10.18653/v1/S19-1031

Sanders, L. M. (1997). Against deliberation. *Political Theory, 25*(3), 347–376. https://doi.org/10.1177/0090591797025003002

Sangiovanni, A. (2017). *Humanity without dignity: Moral equality, respect, and human rights*. Harvard University Press.

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1668–1678). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1163

Sartori, L., & Theodorou, A. (2022). A sociotechnical perspective for the future of AI: Narratives, inequalities, and human control. *Ethics and Information Technology*. https://doi.org/10.1007/s10676-022-09624-3

Schneider, J. (2014). Automated argumentation mining to the rescue? Envisioning argumentation and decision-making support for debates in open online collaboration communities. In *Proceedings of the First Workshop on Argumentation Mining* (pp. 59–63).

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 59–68). ACM. https://doi.org/10.1145/3287560.3287598

Sirrianni, J. W., Liu, X., & Adams, D. (2021). Predicting stance polarity and intensity in cyber argumentation with deep bidirectional transformers. *IEEE Transactions on Computational Social Systems, 8*(3), 655–667. https://doi.org/10.1109/TCSS.2021.3056596

Soral, W., Liu, J., & Bilewicz, M. (2020). Media of contempt: social media consumption predicts normative acceptance of anti-muslim hate speech and islamoprejudice. *International Journal of Conflict and Violence (IJCV), 14*, 1–13. https://doi.org/10.4119/IJCV-3774

Spears, A. K. (2021). African-American language use: Ideology and so-called obscenity. In S. S. Mufwene, J. R. Rickford, G. Bailey, & J. Baugh (Eds.), *African-American English : Structure, history, and use* (pp. 249–276). Routledge.

Springer, N., Engelmann, I., & Pfaffinger, C. (2015). User comments: Motives and inhibitors to write and read. *Information, Communication & Society, 18*(7), 798–815. https://doi.org/10.1080/1369118X.2014.997268

Stahl, B. C., Antoniou, J., Bhalla, N., Brooks, L., Jansen, P., Lindqvist, B., Kirichenko, A., Marchal, S., Rodrigues, R., Santiago, N., Warso, Z., & Wright, D. (2023). A systematic review of artificial intelligence impact assessments. *Artificial Intelligence Review, 56*(11), 12799–12831. https://doi.org/10.1007/s10462-023-10420-8

Stodden, R., Momen, O., & Kallmeyer, L. (2023). *DEPLAIN: A German Parallel Corpus with Intralingual Translations into Plain Language for Sentence and Document Simplification.* https://doi.org/10.48550/arXiv.2305.18939

Stoll, A., Ziegele, M., & Quiring, O. (2020). Detecting impoliteness and incivility in online discussions. *Computational Communication Research, 2*(1), 109–134. https://doi.org/10.5117/CCR2020.1.005.KATH

Stromer-Galley, J. (2007). Measuring deliberation's content: A coding scheme. *Journal of Deliberative Democracy*. https://doi.org/10.16997/jdd.50

Stroud, N. J., Scacco, J. M., Muddiman, A., & Curry, A. L. (2015). Changing deliberative norms on news organizations' Facebook sites. *Journal of Computer-Mediated Communication, 20*(2), 188–203. https://doi.org/10.1111/jcc4.12104

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & Behavior, 7*(3), 321–326. https://doi.org/10.1089/1094931041291295

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., & Wang, W. Y. (2019). Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1630–1640). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1159

Tyagi, A., Uyheng, J., & Carley, K. M. (2020). Affective Polarization in Online Climate Change Discourse on Twitter. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 443–447). IEEE. https://doi.org/10.1109/ASONAM49781.2020.9381419

Uyheng, J., Bellutta, D., & Carley, K. M. (2022). Bots amplify and redirect hate speech in online discourse about racism during the COVID-19 pandemic. *Social Media Society*. https://doi.org/10.1177/20563051221104749

Vecchi, E. M., Falk, N., Jundi, I., & Lapesa, G. (2021). Towards Argument Mining for Social Good: A Survey. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (pp. 1338–1352). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.107

Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0243300

Vrijenhoek, S., Kaya, M., Metoui, N., Möller, J., Odijk, D., & Helberger, N. (2021). Recommenders with a Mission. In F. Scholer, P. Thomas, D. Elsweiler, H. Joho, N. Kando, & C. Smith (Eds.), *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* (pp. 173–183). ACM. https://doi.org/10.1145/3406522.3446019

Weinberg, L. (2022). Rethinking fairness: An interdisciplinary survey of critiques of hegemonic ML fairness approaches. *Journal of Artificial Intelligence Research, 74*, 75–109. https://doi.org/10.1613/jair.1.13196

Wijenayake, P., Silva, D. de, Alahakoon, D., & Kirigeeganage, S. (2020). Automated Detection of Social Roles in Online Communities using Deep Learning. In *Proceedings of the 3rd International Conference on Software Engineering and Information Management* (pp. 63–68). ACM. https://doi.org/10.1145/3378936.3378973

Wojcieszak, M., Thakur, A., Ferreira Gonçalves, J. F., Casas, A., Menchen-Trevino, E., & Boon, M. (2021). Can AI enhance people's support for online moderation and their openness to dissimilar political views? *Journal of Computer-Mediated Communication, 26*(4), 223–243. https://doi.org/10.1093/jcmc/zmab006

Wolfram, W. (2007). Sociolinguistic Folklore in the Study of African American English. *Language and Linguistics Compass, 1*(4), 292–313. https://doi.org/10.1111/j.1749-818X.2007.00016.x

Wyss, D., & Beste, S. (2017). Artificial facilitation: Promoting collective reasoning within asynchronous discussions. *Journal of Information Technology & Politics, 14*(3), 214–231. https://doi.org/10.1080/19331681.2017.1338175

Yang, H., & Callan, J. (2006). Near-duplicate detection by instance-level constrained clustering. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information* (pp. 421–428). https://doi.org/10.1145/1148170.1148243

Young, I. M. (1990). *Justice and the politics of difference*. Princeton University Press.

Young, I. M. (2000). *Inclusion and democracy*. Oxford University Press.

Zerilli, L. M. G. (2014). Against civility: A feminist perspective. In S. Austin (Ed.), *Civility, Legality and Justice in America* (pp. 107–131). Cambridge University Press. https://doi.org/10.1017/cbo9781107479852.005

Ziegele, M., Jost, P., Bormann, M., & Heinbach, D. (2018). Journalistic counter-voices in comment sections: Patterns, determinants, and potential consequences of interactive moderation of uncivil user comments. *Studies in Communication / Media, 7*(4), 525–554. https://doi.org/10.5771/2192-4007-2018-4-525