AI-Driven Mediation Strategies for Audience Depolarisation in Online Debates

Jarod Govers
The University of Melbourne
Melbourne, Australia
jarod.govers@unimelb.edu.au

Eduardo Velloso
The University of Melbourne
Melbourne, Australia
eduardo.velloso@unimelb.edu.au

Vassilis Kostakos
The University of Melbourne
Melbourne, Australia
vassilis.kostakos@unimelb.edu.au

Jorge Goncalves
The University of Melbourne
Melbourne, Australia
jorge.goncalves@unimelb.edu.au

ABSTRACT

Online polarisation can tear the fabric of civility through reinforcing social media's perceptions of division and discord. Social media platforms often rely on content-*moderation* to combat polarisation, contingent on the *reactive* removal or flagging of content. However, this approach often remains agnostic of the underlying debate's ideas and stifles open discourse. In this study, we use prompt-tuned language models to *mediate* social media debates, applying the strategies of the Thomas-Kilmann Conflict Mode Instrument (TKI). We evaluate multiple mediation strategies in providing targeted responses to the debates, as shown to a debate audience. Our findings show that high-cooperativeness TKI strategies offered more persuasive arguments, while an accommodating argument strategy was the most successful at depolarising the audience's opinion. Furthermore, high-cooperativeness strategies also increased the perception that the debaters will reach a consensus. Our work paves the way for scalable and personalised tools that mediate social media debates to encourage depolarisation.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**.

KEYWORDS

social media, debates, mediation, Artificial Intelligence, depolarisation, generative AI, psychology, human-AI cooperation, chatbots

ACM Reference Format:
Jarod Govers, Eduardo Velloso, Vassilis Kostakos, and Jorge Goncalves. 2024. AI-Driven Mediation Strategies for Audience Depolarisation in Online Debates. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3613904.3642322

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642322

1 INTRODUCTION

Social media platforms enable users to discuss ideas, project ideology, and critique political decision-making. From the days of ancient Greece, healthy democracy has been grounded on the marketplace of ideas and in public gathering places for citizens to listen to debates on social and political issues. However, online discussions often become subject to perpetual divisions, mob mentality, and misinformation.

When online debates go awry, a human or virtual moderator may *reactively* intervene to remove toxic or inflammatory posts as dictated by a set of community rules as used on debating platforms (such as self-moderation seen on Kialo and on subreddit communities [69]), or platform-wide community guidelines [23]. However, this form of intervention does not address the conflicting opinions and claims that led to the debate in the first place. It can also shift discussions to more extreme anonymous forums and echo chambers, as seen through the radicalisation of deplatformed (banned) users from mainstream social media such as Facebook, to extreme political fringe groups on Telegram [28, 65] and Gab [16], and political forums such as Stormfront [28, 29].

Likewise, *emotion-driven* content-moderation regulates posts' sentiment and emotion [28, 39, 69]—overlooking that polarisation is a result of *idea-driven* disagreements *that lead to* emotive responses. Thus, fostering constructive online debates with new ideas and questions to the polarised users offers a vital *proactive* means to address polarisation before it can potentially spill into toxic or emotive responses. Likewise, having an objective voice collaborate *in* a discussion is a key measure to help promote consensus and constructive discourse—aiding both the debaters and audience without stifling speech through content-moderation.

Importantly, the victims of polarisation are not exclusively the debaters themselves. In fact, the debaters are often a small minority compared to the thousands or even millions that passively consume social media, sometimes years later [11, 25, 28]. These *audiences and wider communities* are exposed to an increasingly divided debate space on platforms such as Reddit, Twitter, and Facebook [28, 29]. Moreover, exposure to disagreements without solutions or constructive debates can reinforce echo chambers and radicalise users who perceive that societal challenges cannot be resolved (i.e., existential dread) [40, 73]. A lack of understanding and critical thinking within social media debates causes those reading the debates to fall victim

to a polarised 'us-vs-them' mindset, as they perceive that these divisive online debates may never reach a conclusion [31, 79].

Outside of cyberspace, individuals stuck at a polarised impasse can *proactively* discuss their views and concerns with an external mediator—who aim to identify common ground, encourage critical thinking, and propose ideas that both parties may agree to. Encouraging critical thinking and consensus-building can bridge the divide in unresolved online conflicts and encourage an observing audience to consider both sides of the debate with innovative solutions.

However, online debate moderation (i.e., topic-agnostic enforcement of debate rules) and mediation (i.e., voluntary contributions to the debate topic, such as questioning, claims, and proposed solutions which do not rely on removing content) by human users are not scalable given the size and speed of online discussions. The use of AI as a scalable and *proactive* instead of a *reactive* tool in online discourse is increasingly embraced by the HCI community, with research themes such as partisan *debate-bots* and automated argumentation design [61, 63, 71], misinformation in health communication through AI fact-checkers [41], and the psychological perceptions of automated systems in online discourse [10, 41].

Importantly, current approaches do not incorporate psychological conflict resolution theory [2, 6, 75, 88] into the design of AI for conflict-*mediation* online. A mediation-based AI designed around discussion-points, questions, and ideas would encourage further dialogue instead of just deleting controversial posts. This would give users and their audience the choice to either overlook or engage with the ideas and discussion points raised by the mediator-bots. The aim of the mediator's discussion points encourages critical thinking through questioning, solutions, and consensus-driven claims to address social media's culture of antagonism [48, 73] and overcome the 'trench warfare dynamics' of polarised discussions [42].

In this work, we design and test mediation strategies to promote collaborative *idea-driven* mediation on the debate topic over moderating speech itself. Our opt-in mediator-bot design implements the strategies from the Thomas-Kilmann Conflict Mode Instrument—which provides a framework of five conflict resolution strategies based on cooperativeness and assertiveness [44, 75, 76, 88] (i.e., competing/forceful, avoiding/averting, compromising, accommodating, collaborating). Most crucially, our work relies on tuning prompts with the Generative Pretrained Model-4 (GPT-4 [57, 59]) to implement these strategies. These strategies determine how to *question*, *claim*, and *respond* to polarised debaters.

In particular, our study considers the following questions in relation to the strategies we evaluate:

RQ1: How do the strategies influence participants' perceptions of the persuasiveness of the mediator's arguments?

RQ2: How do the strategies influence participants' personal opinion on the debates?

RQ3: Which strategy is the most effective at increasing the audience's perceived likelihood that a debate will reach a consensus?

Our findings show that high-cooperativeness TKI strategies resulted in a higher Perceived Argument Strength score, indicating that these are more effective in providing persuasive arguments to help mediate polarised online debates. We found various degrees of effectiveness across the different conflict resolution strategies regarding how likely they are to depolarise the audience. For example, the *accommodating* strategy was three times more likely to depolarise the audience when compared to having no mediator, while the *forceful* and the *compromising* strategies performed poorly in terms of depolarisation. Moreover, audiences exposed to the *compromising* mediator-bot were 22.2% more likely to believe that the debaters could reach a consensus when compared to having no mediator.

The contributions of our work are three-fold. First, we empirically validate that the psychological TKI conflict resolution strategies can be successfully applied to the *mediation* process of online debates and that a user's preference for a TKI strategy affects the persuasiveness of the strategy. Our manipulation check (Section 3.3.1) also highlights the ability to prompt-tune language models to match specific conflict resolution strategies. Likewise, our psychology-driven GPT-4 prompts are available for users or platforms to experiment with and deploy in their own plug-in mediator-bots for social media. Second, our mediators' ability to depolarise an audience using TKI strategies offers valuable insights for platforms to design non-invasive proactive depolarisation interventions which offer avenues to address divisive debates. Through targeting our mediator-bots on the viewing audience, we can address the larger-scale challenge of community-wide polarisation rather than just the smaller individual-level (i.e., just between debaters) polarisation process. Third, we provide a mixed quantitative-qualitative data-driven discussion on the applicability, design, and limitations of deploying mediator-bots online. Our findings benefit HCI and psychology research by also demonstrating the effectiveness of the TKI strategies in a external *mediator* role rather than the traditional TKI use for active debaters/conflicting person-vs-person engagement.

2 RELATED WORK

To design effective mediator bots, it is important to understand the factors influencing audience depolarisation and perceived consensus-building. In the following sections, we start by defining *mediation*, *moderation* and debate *facilitation*, and their current implementations in HCI research for online conflict resolution. We then examine the psychological conflict resolution theory to elicit the requirements for our mediator-bot experimental design.

2.1 Theoretical Approaches for Resolving Conflict

For the purposes of this study, we consider the following definitions for *moderation*, *mediation*, and *facilitation* in a social media debate context. Firstly, *moderation* relies on overseeing social media content and enforcing a set of rules pertaining to appropriate *conduct*, such as removing hate speech, harassment, or spam, as well as controlling the platform of the debate [23, 74]. A moderator is expected to be topic-agnostic and only intervene when platform policy *breaches* occur—thus being a *reactive* measure for promoting healthy debates.

Similarly, *facilitation* replicates the rules-based order of moderation but relies on encouraging users to participate equally in the debate [39, 74]. A facilitator relies on topic-agnostic rules for ensuring civility (akin to a moderator) but also manages the environment of the debate—such as controlling the conversational tempo

of debaters (to reduce spam) [32, 45], moderating cross-talk [45], and ensuring that debates remain on topic and (if applicable) on time [32, 56]. Importantly, facilitators are not expected to contribute to the content of the debate.

Finally, *mediation* extends debate facilitation by intervening but not overruling a discussion by providing information and *idea*-driven claims, questions, and proposed solutions to promote interaction and degrade out-group mechanics [74]. The objective of a mediator is to encourage cooperation through interactive questioning, providing information/context, and proposing ideas to engage and resolve the debate. Mediators are expected to contribute to the debaters' claims/ideas to help achieve consensus and reduce polarisation [74]. A key distinction between meditation and moderation online is that a mediator is a proactive, voluntary and opt-in process based around a *topic* and addressing *grievances* for constructive discussions. While mediation can be ignored by the audience or debaters,*moderation* is a reactive *response with removal powers* to regulate tone, sentiment, or tempo of a discussion without addressing the underlying ideas [39]—thus it can should only be a 'second line of defence' when mediation fails *and/or* the discussion violates community guidelines/rules.

2.2 Algorithmic Interventions to Address Online Conflict

Existing automated interventions in online debates are limited to debate facilitation, which does not include engaging in the debate itself or proposing solutions/compromises. Conversely, the rise of one-sided *debate chatbots* or social media bots could be exploited as a form of human-targeted polarisation (i.e., influencing a user towards a specific belief) [29, 71, 82]. Meanwhile, existing designs for addressing multiparty privacy conflicts or smart email replies currently rely on wizard-of-oz style experiments between two human interlocutors [35, 66]. Likewise, research in chatbots for interactive ideation includes non-social media tasks such as for cooperative design facilitation [70, 80], thought-mapping to encourage critical thinking [27] and as an aid to help writers create more persuasive arguments [81]. The main challenge of facilitation models is that they do not focus on amending disagreement or reducing polarisation, as debaters may abide by the platform's content rules (enforced by moderators) and debate structure (enforced by facilitators) but still remain in conflict and extremely polarised. Thus, moderators and facilitators alone cannot address the pessimism, narcissism, and mental health challenges of widespread division perceived on social media [11, 48, 73, 78].

Our approach hybridises AI-driven *debating*, *deliberating*, and *consensus-building* to collectively form the basis for debate conflict-mediation [74]. The purpose of the questioning and ideation by a virtual mediator is to invoke perceptions of common ground, problem/debate solvability, and to depolarise the audience. For instance, a mediator could facilitate healthy discourse by identifying and questioning common ground (critical thinking) via persuasive arguments (argumentation strength) and co-design solutions (proposing compromises or solutions).

Importantly, we target persuasive argumentation as the referent object for depolarising the audience, which we measure using the audience's Perceived Argument Strength (PAS) [93]. PAS reflects an aggregate of nine questions into a 1.0-to-5.0 score to reflect an arguments ability to influence user opinion and behaviour—with moderate replicability and reliability across psychological studies [4, 37, 38, 93]. In HCI literature, Karinshak et al. [41] identified that using AI-generated pro-vaccination messages resulted in higher Perceived Argument Strength (PAS), and thus more persuasive arguments [93], compared to those created by the United States Centre of Disease Control.

2.3 Conflict Theory—The Groundwork for Mediation

To create a theory-driven framework for mediation bots, it is essential to understand the psychological strategies that humans employ to resolve conflict. In organisational behavioural management theory, Blake and Mouton's managerial grid classifies human conflict into the axis of *concern for people* and *concern for production* [6]. It posits that individuals balance risks (such as the risk of escalating the conflict, making enemies, or 'losing' the debate) with potential social or tangible rewards—such as romantic interests, economic gains; collectively also known as *Social Exchange Theory* [36].

Psychological measures for social exchange theory are measured through questionnaires to identify a participant's preferred style to resolve conflict. The standard in psychological research is the Thomas-Kilmann Conflict Mode Instrument [44]—which extends on the tangible (material gains) vs. intangible (social and metaphysical) cost-benefit interaction effect through a five-strategy conflict instrument with a *cooperativeness* and *assertiveness* axis.

It is important to note that all strategies aim to *cooperate* with both parties to come to an agreement. In the context of the TKI axis, the definitions of *cooperativeness* and *assertiveness* are:

- *Cooperativeness*—the degree of which a party attempts to address the specific grievances of the opposing party(s) [75]. Low-cooperative strategies (Competing, Avoiding) focus on the 'wider picture' and utmost over targeting smaller individual problems/disagreements.
- *Assertiveness*—the degree of which a party pushes their own concerns [75] (in this case, the Mediator's confidence and directness in its questioning and claims to achieve consensus).

For instance, it is not trivial that high-cooperativeness strategies' will result better conflict resolution outcomes [44, 88]. In a debate setting, high cooperativeness would consist of highlighting and questioning *specific* grievances—thus *expanding* the debate and opening more room for criticism which could be seen by a viewing audience as complicating the debate. Likewise, high-cooperation can be perceived as overthinking and lead to analysis paralysis where decisions get stuck on specific inter-party grievance [75, 76, 88]. Conversely, low-cooperativeness approaches such as *forcing* a solution, diverting (or *avoiding*) the issues for a less controversial approach, and a neutral/'middle-ground' compromise are valid resolution approaches. Hence, our experiment aims to investigate the effects of *cooperativeness* and how confident/*assertive* the mediator is in intervening in the debate to persuade the audience towards an amicable debate outcome.

The strategies represent each corner of the TKI as well as the central 'Compromising' strategy [88], as visualised in Figure 1. These strategies are typically defined as follows:

- Collaborating—assertive and cooperative, mutual problem-solving to satisfy both parties' needs;
- Compromising—intermediate in both assertiveness and co-operation, exchanges concessions;
- Competing—assertive and uncooperative, tries to win own positions;
- Accommodating—unassertive and cooperative, satisfies the other's goals;
- Avoiding—unassertive and uncooperative, postpones or avoids unpleasant issues.

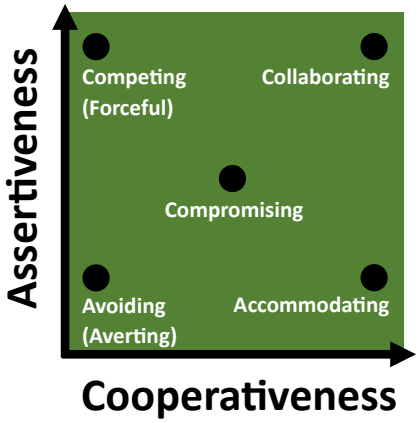

Figure 1: The TKI Assertiveness and Cooperativeness axis quadrant, visualising the five conflict resolution strategies.

Alternate measures for identifying a participant's preference of resolving conflict include those by Lawrence and Lorsh [49] or Hall [33]—which also derive from the Blake-Mouton grid but differ on collection type (likert vs binary choices) and question categories. However, TKI offers more consistent results when a participant repeats the measure to (re)identify their preferred strategy (known as *test-retest reliability* [15]; with the highest performing 0.64 mode score for TKI [76] compared to 0.55 for Hall [33], 0.50 for Lawrence-Lorsh [49], 0.39 for Blake-Mouton [6]). Likewise, the TKI assessment questions are more likely to reflect the desired strategies, improving our confidence when making claims regarding a user's preference for a conflict-resolution strategy—as measured by Cronbach alpha metric for internal consistency (higher is better [15]; with 0.60 for TKI [76], 0.45 for Lawrence-Lorsh [49], 0.55 for Hall [33]). As such, TKI's consistent results when repeated on the same individual at different times (i.e., *high test-retest reliability*), alongside TKI's relevant questions that match the psychological theory (*internal consistency*) forms the basis for why we utilise the TKI strategies for our mediator-bot design.

Building on Thomas and Kilmann's work, Womack identified that users prefer solutions to conflicts based on one of the five TKI strategies [88]. Hence, we hypothesise that a mediator who utilises the participant's preferred conflict resolution strategy will perceive the matching mediator as more influential in reducing the difference in agreement between two debaters (i.e., polarisation) and increase the participant's belief that the debaters would reach a consensus.

The TKI measure does not assume neutrality in all strategies [75], nor does mediation itself [74]. Thus, we do not require or force the mediator-bot to be 'neutral' or adopt a middle position. Only, the 'Compromising' strategy represents a middle-ground 'give-and-take' approach by design.

2.3.1 Addressing conflicting debaters through audience perceptions. Polarised online debaters aim to defend and protect their own group's identity (in-group) against those with opposing views (i.e., the out-group) [58]—with Das and Kramer identifying that the presence of a third perspective between diametrically opposed debating parties reduced the aggressive tone and personalised attacks towards the opposing party in Facebook discussions [18]. As such, a growing body of research highlights the role of the *audience* in influencing debater behaviour to reduce hostility [18, 42, 50, 72]. Su et al. identified that social media users engage in one-sided speech when they are aware that their target viewing audience also hold one-sided beliefs [72]. Marwick and Boyd also identified that social media users placate their narratives and ideology around an 'imagined audience' consisting of all parties engaging in the debate as well as the perceptions of the wider viewing audience [55].

Finally, the perception of the audience is highly relevant to online debates as any future post will be from either a past audience member or the original debaters—who seek to appease their audience. Thus, this symbiotic reader-writer relationship is pertinent to building a culture of open, critical and constructive debate.

3 METHODOLOGY

The key objective of this study is to utilise mediator strategies through virtual agents to reduce a social media reader's (the audience) *personal* opinion polarisation on polarised online discussions, as well as reduce their *perception* that the debaters are polarised. We expect the strategy-driven mediator-bots to provoke constructive questions and solutions and invoke audience introspection on the debates—as these set the requirements for critical thinking and changes in opinion required for depolarisation [90].

We operationalise this objective by measuring the audience's argument strength rating for each mediator (RQ1), as well as their change in personal agreement on the topic after reading the debate transcript (with or without the mediator-bot, for RQ2), and their perceived belief that the debate would reach a consensus given the debaters' polarised views (RQ3).

We grounded our mediator bots in social psychology research, tailoring different mediator bots to model the five TKI strategies. In the following sections, we highlight the design considerations for creating the debate transcripts and the mediator bots and report the results of a manipulation check to verify that the prompts used for our mediator bots match their assigned TKI strategy across the debate topics.

3.1 Debate Transcript Design

We drew the debate text from Kialo—an open debating platform with over 2 million user contributions across over 18,000 public debates [43]. The academic benefit of Kialo is its bilateral nature, whereby users debate a specific debate topic with a specified "For" or "Against" stance—thereby allowing annotated and definitive polarised stances. Kialo also enables users to create reply-chain

threads where users can reply and argue for or against statements. Thus, these reply threads are beneficial as we can replicate the *conversational back-and-forth* nature of online arguments while maintaining realistic user stances.

As the key focus of the mediator-bots is to voluntarily help remediate the *ideas/concepts* leading to the increase in social media division/polarity relevant for content-*mediation*, we do not target the *emotive-driven* reasons for polarisation—as this is for the separate domain of content-moderation. As such, our *idea*-driven *mediation* over *emotive*-driven *moderation* approach sees our mediator-bots as a first-line of defence in the assumption of genuine debate, while personal flame-wars and heated attacks requires a separate 'second-line-of-defence' *emotive-driven* and topic-agnostic approach. The field of *reactive* regulation of tone, sentiment, and tempo is already established in prior content-*moderation* work, and thus out-of-scope for our novel *first-line proactive* mediation approach [39]. Thus, Kialo (a platform with content-moderation guidelines similar to Reddit) offers a compelling ideas-driven baseline 'ground-truth' for our mediator study—while still containing controversial and personal topics such as COVID-19 vaccine mandates and the Russia-Ukraine war.

As our intervention group tests each of the five TKI conflict resolution strategies, we collected five debates covering social, economic, and cultural topics to represent the variety of topics seen in online debates. Specifically, the chosen debate topics were:

(1) "Public transport should be free." (PT)
(2) "COVID-19 vaccines should be mandatory." (CV)
(3) "Democracies should take in both Ukrainian refugees and Russians seeking to escape conscription in the Russia-Ukraine War." (UKR)
(4) "Facebook should ban political ads with misinformation." (FB)
(5) "US intelligence agencies should stop mass data collection." (NSA)

We selected the first top-level response to the debate topic on the condition that it has a reply that is *for* the statement and a reply that is *against* the statement to ensure that the debate remains balanced with one parent claim and a reply from both sides. We also selected this statement-reply structure to test the adaptability of our mediator bots to inter-personal conflict/arguments, rather than a non-conversational 'list of statements for and against' as seen with competing debate platforms such as ProCon [62].

3.2 Bias Mitigation—Debate Transcript Visual Design

We selected Reddit as our design reference for a *mainstream* discussion platform to emulate due to its familiarity across social media users (unlike Kialo) and its polarised communities [11, 28, 31], as well as the structure of Reddit's conversations (i.e., grouped topic threads, reply threads).

We made the following design choices in our Reddit-style visual design to mitigate potential biases, as also represented in Figure 2:

- *Debater usernames*: Each topic transcript uses a different username for the pro and anti-debate topic users (e.g., Person A|B for the COVID-19 topic pro|anti debaters, Person C|D for the Public Transport topic, etc.) to remove any persistent biases from prior arguments.
- *Colour-choice*: Person A|B, C|D, etc. all utilise opposing colours, with pure red and green removed from the gamut due to their connotations with good or bad [1].
- *Removed votes/Karma*: no comments have any votes/karma or other Reddit tags to avoid conformity biases.
- *Reply order*: Each debater has one main post each, with two replies—one from the (anti-topic) user and another from their (pro) position. The ordering is fixed across all topics and interventions.
- *Timing*—to avoid the perception that the debaters are 'ignoring' the simulated mediator, we indicate that all mediator posts were sent '1 minute' ago from the fake Reddit transcript 'screenshot'.
- *Mediator-bot username*: We do not anthropomorphise nor offer a gendered name for the mediator bots [3, 83, 84, 87]. Framing our intervention as a 'Mediator-bot' ensures external validity by ensuring that the audience is not deceived into believing the mediator is real—as a real-world use case would see the bots deployed as a collaborative AI tool. Likewise, naming the bot based on <FUNCTION-BOT> emphasises the role and virtual nature akin to prior virtual agents (e.g., FacilitatorBot [39], TaskBot [77], Debbie the Debate-bot [63]).
- *Subreddit layout*: the choice of subreddit name does not reflect any major debating subreddits to avoid community bias and stereotypes [11].
- *Omission of reference links to data sources:* the Microsoft GPT-4 adaptation used in this study collects live information from the internet to contextualise and present factual information in its responses [57]. GPT-4's sources are provided as citation links (e.g., 'Sources: [1] author, source, date' format). We removed references to the original link in the output citations to avoid biases surrounding the source's validity.
- *Reply length:* we control for the length variance of GPT-4 responses to ensure the overall mediator reply word count are balanced between strategies (+- 20% word count variance between speech acts) per topic. We employ a soft-approximate approach by instructing GPT-4 to produce a short paragraph per speech act rather than a hard word-count limit to avoid abruptly cutting off or stifling GPT-4's responses.

Figure 2 displays a debate topic transcript example from the intervention group for the Russia-Ukraine refugees (UKR) topic with the mediator using the Collaborating TKI strategy.

3.3 Prompt-Engineering—Designing a Standard Mediator Prompt Template

We used GPT-4 to generate our mediator-bot text due to its better performance compared to other LLMs on reading comprehension (80.9% F1-score on the DROP reading comprehension and arithmetic benchmark [21, 59]), and the HellaSwag deductive reasoning benchmark (95.3% with 10-shot learning [59, 92]).

We used Microsoft Bing's GPT-4 implementation with its online feature to automatically collect new information and context for its responses [57]—ensuring relevance given recent events such as the 2022 Russian invasion of Ukraine, which was not captured in the original ChatGPT/GPT-4's September 2021 corpus [59]. Figure 2 displays an example of the mediator's response and knowledge.

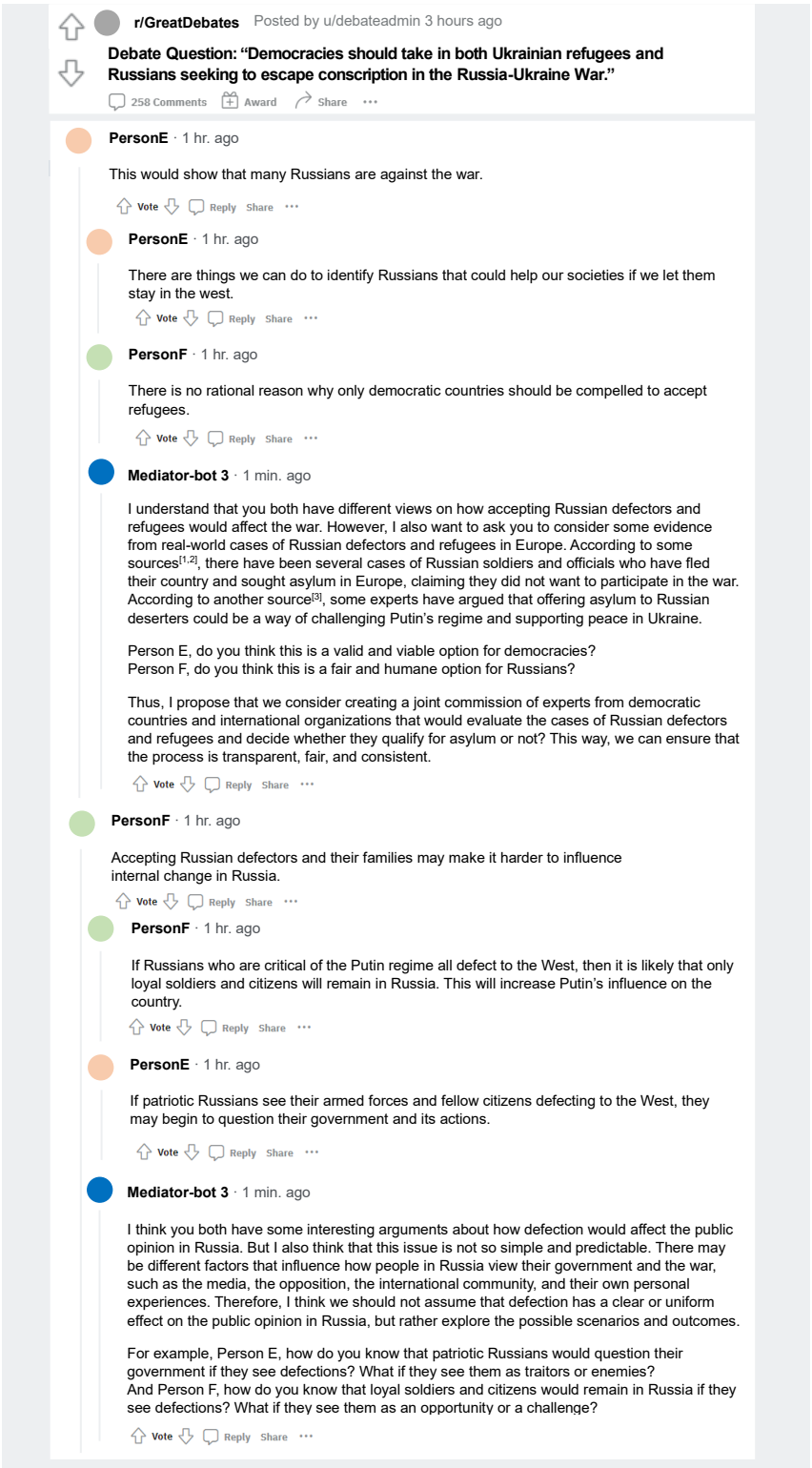

Figure 2: Layout for the Russia-Ukraine debate topic (UKR), displaying the pro-stance (Person E) and anti-stance (Person F) debates, alongside the in-depth (high-cooperativeness, high-assertiveness) Collaborating strategy mediator.

We followed an iterative and incremental prompt-engineering process to ensure the mediator replicated only one of the five TKI strategies. Throughout this process, the TKI strategy 'Competing' was renamed to 'Forceful' to reflect that the mediator themselves is not competing to the debate itself, but *forcing* its own views into the debate. Likewise, the 'Avoiding' strategy was renamed to 'Averting' to reflect that the mediator will always remain present in the debate and not *virtually* leave/not-respond, but rather try to avert the debate question and instead opt for less controversial solutions as reflected in its TKI definition. All GPT-4 prompts were topic-agnostic and were not provided with any contextual information on the topic, with our prompts available in our supplementary material. Our prompts consisted of four paragraphs:

(1) A paragraph explaining each TKI strategy from the mediator's perspective. We also use these definitions to help with the manipulation check to ensure that independent raters can identify which mediator bot output corresponds to the TKI strategy.
(2) The text of the debate, with replies provided as 'Person A, Statement 1: ...', 'Person B replying to Statement 1: ...'.
(3) Response speech act options—consisting of *Ask Question*, *Make Claim*, and *Respond*; this design choice ensures a consistent output structure from GPT-4 and is based on the predicate-logic *debate speech acts* framework proposed by Prakken [61]. All speech acts are equally balanced across the debate transcripts and Mediator strategies.
(4) A generic topic-agnostic instruction prompt for GPT-4 to role-play as a mediator whose aim is to achieve consensus while not using the specific names of the TKI strategies to mitigate bias for those aware of the TKI strategies.

3.3.1 *Manipulation Check.*

To ensure that our prompts accurately operationalise the TKI strategies, we relied on independent raters to read the mediator bot responses and group them by the TKI strategy. Specifically, we utilised the Prolific crowdsourcing platform to recruit 20 participants who were based in the United States to ensure the relevance of the current events topics; were native English speakers, to mitigate challenges with language proficiency [32]; and had approval ratings above 97%. All participants were compensated based on the highest minimum wage available across all states of the United States ($15.74USD) at the time of data collection, which is well above the payment recommended by Prolific [20].

Through the Qualtrics survey platform, raters viewed the five debate transcripts along with all mediator bot responses representing each of the five TKI strategies. They then matched each mediator to their perceived TKI strategy. The debate topic and order of strategies presented were both randomised to mitigate bias.

We measured the level of agreement through the accuracy of the raters' classifications and via Fleiss' Kappa for inter-rater agreement [26]. We visualise our raters' classifications in Figure 3, where we contrast their predictions with the actual strategy the mediator was trying to convey. In general, we obtained 88.21% matching classifications, indicating that raters were able to identify the intended strategy in most cases. In addition, we attained a mean Fleiss' Kappa of 0.73 (as categorised in Table 1), whereby a Fleiss' Kappa between 0.61 to 0.80 indicates 'substantial agreement' and 0.80-1.00 indicates 'almost perfect' agreement between all raters [26].

Table 1: Fleiss' Kappa agreement for each one of the debate topics.

Debate Topic	Fleiss' Kappa
PT	0.83
CV	0.70
UKR	0.75
FB	0.65
NSA	0.72
Overall	**0.73**

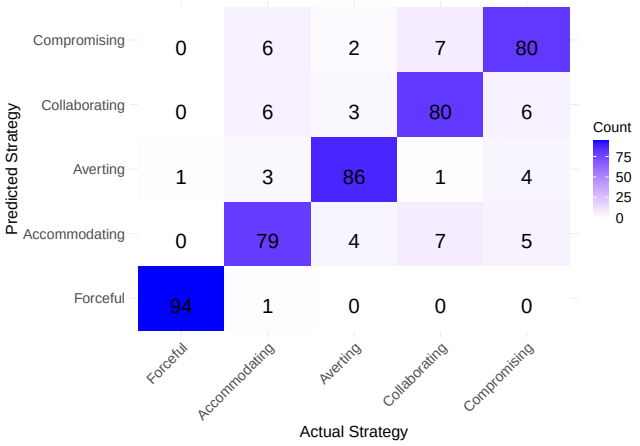

Figure 3: TKI strategy classifications' confusion matrix from the manipulation check.

Given our results, we can confidently conclude that our topic-agnostic prompts and their responses effectively reflect the mediator-bots targeted TKI strategy.

3.4 Procedure

We recruited a total of 144 participants with an average age of 39.06 (SD = 13.29) years old using the Prolific platform, applying the same pay rate and filtering conditions as stated in our manipulation check, with the addition of equal-split balancing for political party preference (Republican, Democrat, Independent). The sample size was based on a G* Power analysis using a f^2 effect size of 0.15 (indicating a medium-effect size as used in similar psychological intervention studies [14, 47, 53, 68]).

Our survey consisted of several sections and took approximately 45 minutes to complete. Figure 4 outlines our experimental workflow. Our experimental design was approved by our university's Human Ethics Committee.

Stage A started by providing the participant with a plain language statement, explaining the details of the experiment. This was followed by a 'Pre-task Questionnaire', in which we collected participant information, including demographic data (age, gender and

education level) and their prior online debate experience through two 5-pt Likert scales ('Never', 'Sometimes over the past six months', 'Sometimes over the past month', 'Weekly', 'Daily'). Participants were asked *"How often do you read/watch (without posting) debates on social media?"* for reading/*passive experience*, and *"How often do you post and actively partake in debates on social media?"* for writing/*active engagement*. Participants were also asked to complete the Adkins Self-Questionnaire, containing 15 questions across the five strategies to categorise one's preference for TKI assertiveness and cooperativeness in relation to the five strategies [2]. This tested if TKI preference enhances the corresponding mediator strategy's effectiveness. We utilise the normalised ratio of the participant's preference for *assertiveness* and *cooperativeness* as the two TKI preference indicators—representing where they stand on the TKI assert/coop-axis quadrant from Figure 1. We also collected the participants' pre-existing topic agreement on each of the five debate topics using a 7-point Likert scale ('Strongly Disagree' to 'Strongly Agree').

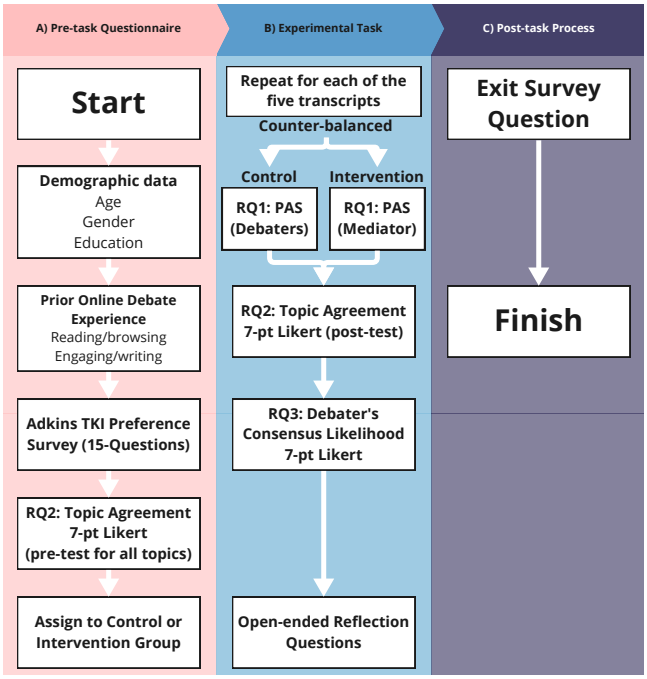

Figure 4: The experimental flow of our study—covering demographic, predictor and pre-test questions in Stage A, the iterative per-transcript questions across the separate Intervention and Control groups in Stage B, and the post-task exit question in Stage C.

Next, participants were assigned to either the control group (without any mediators) or the intervention group (with one of the five mediator strategies per transcript, counterbalanced). Stage B reflects the iterative process whereby users read the different debate transcripts. After reading each transcript, participants were asked to answer the following questions covering each of our research questions:

RQ1: How do the strategies influence participants' perceptions of the persuasiveness of the mediator's arguments?
- **Quantitative Measure: Perceived Argument Strength (PAS) score** for the Mediator (intervention-group) and Debaters (control-group, for reference only)—a numeric score between 1.0 to 5.0 based on 9 Likert questions. PAS provides an objective means of gauging the impact of different rhetorical strategies and persuasive techniques and enables the evaluation and refinement of counterarguments, as it enables researchers to pinpoint the weaknesses in opposing viewpoints as perceived by the target audience. PAS demonstrates high replicability [4, 37, 93], unlike measuring argument strength as a binary (strong vs weak) option or as a single Likert scale [37].
- **Qualitative Measures:** An open-ended question on the mediator's approach towards mediating the debate: *"What were your thoughts on the mediator's strategy to mediate in this debate?"*.

RQ2: How do the strategies influence participants' personal opinion on the debates?
- **Quantitative Measure: Topic Agreement Score**— comparing Pre-test debate topic agreement vs. Post-test debate topic agreement 7-pt Likert between 'Strongly Disagree' to 'Strongly Agree'. Namely, we measure the audience's agreement on the debate transcript's question (such as "Public transport should be free") prior to showing the debate transcript and after they completed steps within that particular transcript.
- **Qualitative Measure:** An open-ended question on the participant's introspection and reflection on the debate to investigate the logic and causes between their shift (or lack thereof) of opinion after reading the debate: *"Did the mediator help you reflect on this debate's topic? If so, in what way?"*

RQ3: Which strategy is the most effective at increasing the audience's perceived likelihood that a debate will reach a consensus?
- **Quantitative Measure: Perceived Consensus Score**—a 7-pt Likert between 'Extremely Unlikely' to 'Extremely Likely' on *"How likely do you think that Person A and Person B would come to an agreement or solution to the debate?"*
- **Qualitative Measure:** An open ended response to *"Explain why you chose your prior answer."*

After iterating over all five debate topic transcripts in their respective group (control or intervention), participants were asked a final exit survey question (Stage C) on their opinion on *"What are your thoughts on deploying virtual bots to mediate debates in social media?"*. This was followed by a message thanking participants for completing our experiment.

4 RESULTS

We employ a mixed-methods approach with linear mixed models and thematic analysis to address each one of our research questions. We utilise a Generalised Linear Mixed Model (GLMM) to test the PAS values across the different strategies which account for our predictor values (RQ1). We utilise Cumulative Link Models (CLM)

to represent the ordinal Likert data across the mediator strategies, targeting personal topic opinion polarisation (RQ2) and perceived debater consensus/polarisation (RQ3). We consider the mediator strategies in both their one-of-five strategy form (*Accommodating*, *Collaborating*, *Compromising*, *Averting*, *Forceful*), and in their high-medium-low *Assertiveness* and *Cooperativeness* axis form, which we infer from the TKI quadrant visualised in Figure 1. We also consider the participant's TKI preference from the Adkins TKI self-questionnaire—where we utilise their preferences towards the cooperativeness and/or assertiveness axis.

We measure the effect size of the strategies and relevant predictors using unstandardised coefficients for PAS response and Likert scale data, alongside Odds-Ratio (OR) and Cohen's d standardised effect sizes for binomial data (such as whether the mediator depolarised the audience participant or not) [14]. For pairwise comparisons, such as comparing the mediator TKI strategies, we utilise Estimated Marginal Means comparisons (i.e., *emmeans*, also known as *least-squares means*). Emmeans are useful for representing the wider population as they account for the effects of each variable/predictor in our models as opposed to the adjusted raw sample means. For significance, we provide p-values and Standard Error (SE) for each effect-size claim. We compute the Variance Inflation Factors (VIF) to check for multicollinearity across our independent variables for each mixed model. All VIFs for our models were below 5, indicating a lack of linear dependency among the independent variables [64]. Our model outputs (all generated in R) and final topic-agnostic GPT-4 prompts for each strategy are available in our supplementary material.

Finally, we conducted a deductive thematic analysis of our qualitative responses, which involves coding them into themes based on argument strength (RQ1), critical thinking (RQ2), and conflict-resolution (RQ3) theories. Deductive thematic analysis involves an iterative process of familiarising frequent concepts in the data, grouping trends as themes, reviewing themes with the theory [7]. We focus our theory-based analysis on Critical Thinking and speech act models [61, 89, 90], and the psychology of the TKI strategies [44]). We iterate over the data across three passes and present the relevant information for each research question.

4.1 RQ1—Perceived Argument Strength by TKI Strategy

Table 2 shows the mean PAS for each strategy and all strategies combined. It also shows the mean PAS of the Human Debaters as rated by our participants in the Control condition (no mediator) as a reference.

We then tested the difference in Perceived Argument Strength values (1.0 to 5.0) across the five mediators from the 120-person intervention group and 24 control group (with the debaters argument strength as the measure). We found a statistically significant relationship between the mediator strategy and its perceived argument strength—whereby Collaborating, Compromising, and Accommodating (i.e., the more *cooperative* strategies) outperformed the Forceful and Averting strategies (β = 0.38, SE = 0.07, p < 0.05). Likewise, Collaborating and Compromising (i.e., the cooperative *and* assertive strategies) had statistically significant PAS improvements over the control debaters' argument strengths (up to β = 0.45, SE = 0.12, p < 0.01). All pairwise strategy comparisons are outlined in Table 3. Figure 5 highlights the distribution of PAS scores given to each strategy. The full fixed-effects model output is available in the supplementary materials ($\mathbf{R}^2$: 0.39).

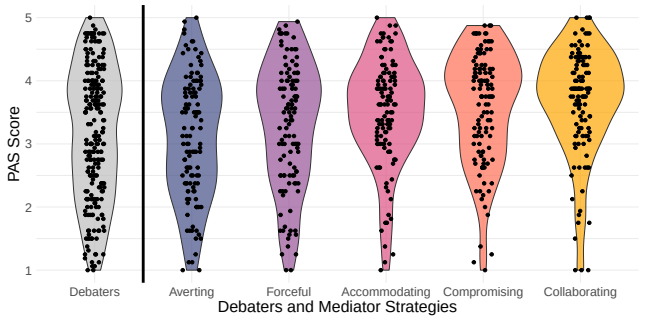

Figure 5: Distribution of Perceived Argument Strength Scores between the Debaters (Control) and the Mediator Strategies.

Table 2: Descriptive statistics on PAS scores for the different mediator strategies, and human debater PAS scores (from all debate topic transcripts in the control condition) for reference.

Strategy	Mean PAS	Standard Deviation (SD)
Averting	**3.113**	**0.970**
Forceful	**3.299**	**0.988**
Accommodating	**3.493**	**0.857**
Compromising	**3.579**	**0.883**
Collaborating	**3.667**	**0.856**
Avg. All Strategies	**3.430**	**0.931**
Avg. of Human Debaters (Control)	**3.156**	**1.049**

4.1.1 *Effect of TKI Preference on PAS.*

Prior work has highlighted that people prefer resolving conflicts linked to their preferred TKI strategy [2, 44, 67, 75, 76, 88]. Hence, we use our GLMM model to cross-analyse each participant's preferred TKI strategy and compare their assertiveness and cooperative scores to the strategy used by the mediators.

We observed a statistically significant positive relationship between a user's preference for more cooperative strategies and PAS scores for the medium-cooperativeness Compromising (β = 0.102, SE = 0.036, p < 0.01) and high-cooperativeness Collaborating (β = 0.074, SE = 0.036, p < 0.05) mediators. We also observe a statistically significant and moderate increase between preference for more assertive mediation strategies and PAS scores for the medium-assertiveness Compromising strategy (β = 0.177, SE = 0.080, p < 0.05).

Conversely, Forceful and Averting lacked a statistically significant relationship between preference for these high-assertiveness or low-cooperativeness strategies and PAS scores (β = 0.02, SE = 0.04, p = 0.63 and β = 0.07, SE = 0.08, p = 0.41).

Table 3: Pairwise comparison of PAS coefficients across mediator strategies (Model A) and across the assert/coop TKI axis (Model B).

Contrast (across TKI *strategies*)	Estimate	Standard Error (SE)	p-value
Forceful - Accommodating	-0.200	0.094	0.279
Forceful - Averting	0.186	0.094	0.356
Forceful - Collaborating	**-0.372**	**0.094**	**0.001**
Forceful - Compromising	**-0.293**	**0.094**	**0.023**
Forceful - Control	0.082	0.115	0.980
Accommodating - Averting	**0.386**	**0.094**	**0.001**
Accommodating - Collaborating	0.172	0.095	0.456
Accommodating - Compromising	-0.094	0.094	0.920
Accommodating - Control	0.282	0.114	0.137
Averting - Collaborating	**-0.558**	**0.094**	**<.001**
Averting - Compromising	**-0.480**	**0.094**	**<.001**
Averting - Control	-0.104	0.115	0.944
Collaborating - Compromising	0.078	0.094	0.962
Collaborating - Control	**0.454**	**0.115**	**0.001**
Compromising - Control	**0.375**	**0.114**	**0.01**
Contrast (across the TKI *axis*)	**Estimate**	**Standard Error (SE)**	**p-value**
High-cooperativeness - Low-cooperativeness	**0.378**	**0.069**	**<.001**
High-assertiveness - Low-assertiveness	**0.177**	**0.069**	**0.011**

4.1.2 Perceptions of the Mediator's Approach.

Stronger arguments had a lasting psychological impact, with the more cooperative strategies promoting critical thinking—as imbued in statements when asked whether the mediator made them reflect on the debate; *"Absolutely, yes. The approach allowed me to keep an open mind and try to figure out solutions, rather than solely defend my opinion/my side"* (P67) and that *"The moderator brings a way for both parties to understand the pros and cons of each other's statements"* (P31).

Participants identified the high cooperativeness and assertiveness collaborating mediator as quite successful in providing persuasive arguments (as also shown in our quantitative findings), with statements like; *"The [collaborating] mediator followed an effective, fair, and formulaic strategy: acknowledge both sides, prompt for sources/further information, provide a diplomatic solution. I believe that this has been the mediator's best approach/strategy thus far"* (P67). The high-assertiveness and low-cooperativeness forceful mediator was far more divisive. Some participants perceived it as *"rational"* (P2) and *''direct"* (P17, P21, P49) in its approach to win over the debaters to a specific side with claims such as *"I enjoyed the way it called out person F!"* (P1) and *"It called out both people on this topic which I enjoyed"* (P99). However, others criticise the mediator's *"aggressive"* (P60, P81, P112, P118) approach in that it did not seem receptive to discourse; *"right out of the gate turned me off and made me shut down"* (P52) and that the Forceful mediator appeared *"combative and contrarian"* (P19).

Regarding the *tone* of the mediator, some participants found that low-assertiveness strategies had a polite depolarising effect on the perception of the debaters through *casualising* the discussion—imbued through claims such as *"I think that the mediator was exactly correct here, in trying to reroute the discussion to something that was actually on topic and not needlessly inflammatory and judgmental"* (P18). Likewise, the casual/polite persona and recommended (but not forced) solutions enhanced the perception that the *debaters* were not polarised; *"The bot was good at proposing middle ground solutions"* (P51).

However, while low-assertiveness resulted in the perception of *"calm[ing] things down"* (P98) for some participants, its arguments attained the lowest PAS values in our quantitative analysis. Several participants echoed this degradation of PAS due to the mediator's low assertiveness as they felt these mediators appeared to lack *confidence* and saw the debate as *unsolvable*. Participants expected the mediator to take a more authoritative role to be more than just another debater *"in which it's trying to act more human and cavalier in order to be more accepted. It's a technique that generally backfires pretty badly"* (P74) and that the mediator *"sounds slightly unprofessional, and should have provided sources instead of saying "I read some things" etc."* (P17). Furthermore, they perceived the informality of an unassertive casual approach as a negative trait of the mediator-bot as *"It detracted from the ideas the mediator put forth"* (P70). In some situations, the low-assertiveness strategies' *"politeness"* (P57, P21, P2, P89) was even misconstrued as *passive-aggression* which *"came across as pretentious and unproductive."* (P25) and that the *"mediator's tone was weak and too soft, which made it seem less credible."* (P67)—thus showing a desire for more direct argumentation.

4.2 RQ2—Change in Audience's Personal Opinion via the Mediator-bots

Next, we evaluated participants' change in their own *personal* opinions about the debate topics. We present our full model in the supplementary material. Table 4 shows a statistically significant

relationship between the mediator strategy and whether their post-test (i.e., after reading the transcript) topic agreement Likert score differed from their original pre-test topic agreement Likert score (i.e., before reading any debate transcripts). We observed that strategies that lean towards the *cooperative* TKI axis (i.e., Accommodating, Collaborating, and Compromising) had a statistically significant influence in changing their opinion compared to the control. The highest-performing Accommodating strategy was 2.8 times more likely to change their opinion compared to the Control (OR = 2.817, Cohen d = 0.571, $p < 0.01$; as derived from Table 4). Our model also highlights that users who read or write in online debates 'sometimes over the past month' are approximately half as likely to change their mind as those who read online debates 'daily' (OR = 1.584, Cohen d = 0.254, $p < 0.05$ for reading; OR = 1.917, Cohen d = 0.359, $p < 0.05$ for writing).

Beyond binary opinion changes, we also categorise if the audience were depolarised, as defined by the difference from the neutral stance in the 7-pt topic agreement Likert scale between the participant's pre- and post-test topic agreement value. Thus, those moving towards the centre after our manipulation were considered as *depolarised* while those that moved away from the centre were considered as *polarised*.

4.2.1 Effect of TKI Preference on Audience Opinion Depolarisation. Users who preferred cooperative strategies had a slight increase in the likelihood of being depolarised when subject to the Compromising mediator (OR = 1.460, $p < 0.05$) with a small effect size for TKI preference in audience opinion depolarisation (Cohen d = 0.209).

Figure 6 displays the movements of user opinion, denoting the (de)polarisation process. We observe that participants were approximately three times more likely to be depolarised from their original stance when exposed to the Accommodating mediator compared to the no-mediator Control debate transcript (OR = 3.001, Cohen d = 0.606, $p < 0.05$). Interestingly, the Compromising mediator had a somewhat blowback dividing effect, being the strategy that led to the highest number of polarised participants. In addition, Averting's focus on deflecting the debate towards less controversial areas/topics also had notable depolarisation effect despite its low-cooperativeness/low-assertiveness.

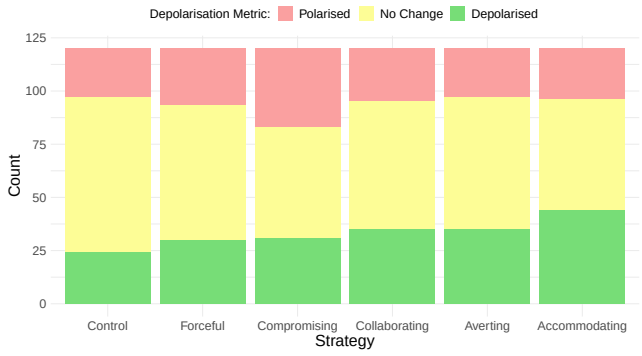

Figure 6: Audience change in opinion between strategies and the no-mediator Control group, as categorised between being polarised, depolarised, or neither/no change.

4.2.2 Critical Thinking and Change in Opinion. Though the audience was not active in the debate, they evaluated and raised questions regarding the mediator's role in helping them engage in critical thought. The Collaborative mediator encouraged introspection even in cases where participants did not change their original opinion, with participants claiming that *"...the bot made me think more about where I actually stand on this. I think I need to do some more searching because free speech is important to me"* (P81), and *that "It [the mediator] did help me reflect, but caused me to feel stuck on deciding one way or the other. I can't choose now because I am trying to weigh each side."* (P119). Likewise, participants stated that they tended to respond more to this mediator's ideas with comments like *"... it allowed me to keep an open mind and try to figure out solutions, rather than solely defend my opinion/my side"* (P67) and that it used a *"strategy of encouraging participants to question their beliefs and explore the topic deeper [which] is a proactive approach to facilitating meaningful discussion and potential resolution"* (P107).

Conversely, the high-assertiveness but low-cooperativeness Forceful strategy made participants engage in *negative* information denial and analysis—whereby participants sought to find reasons to *invalidate* arguments. This mediator had an effect of further polarising preexisting beliefs, with claims that it made participants *"double-down"* (P86, P39, P73, P84, P89) on their view, and could also aggravate them to feel *"more inclined to respond with hostility to those that hold opposing beliefs"* (P89).

Interestingly, despite Accommodating and Compromising both being *cooperative* strategies, Compromising had an unexpected polarising effect. Participants either embraced or rejected the mediator's applicability and realism, with claims that a *"quick and easy solution to a complicated topic like mass surveillance is hilariously naive."* (P110), thus its *compromise*-based strategy was insufficient for eliciting critical thought in nuanced discussions. Meanwhile, Accommodating's polite and unassertive questioning *"validated their points"* (P66) and made the participants feel that the mediator had a use in building *"civility in our discussions but to also combat the rise of misinformation and propaganda"* (P29).

Surprisingly, the low-assertiveness / low-cooperativeness Averting mediator was mildly successful with regard to depolarising participants. The participants perceived the Averting mediator as diffusing hostility by proposing alternative, less controversial solutions, questioning the relevance of the debate topic and attempting to adjourn it. For instance, users claimed that Averting was *"...an effective strategy for sensitive topics especially"* (P67). However, the audience contested Averting's effectiveness as a mediation strategy due to its nature to *avoid* resolving the specific issue, with users claiming that this approach was *"counterproductive... [as] arguing can be important sometimes and this disagreement [NSA surveillance debate] was happening with civility, with people just explaining their opinions - no name calling or personal attacks."* (P118).

4.3 RQ3—Perceived Likelihood that a Debate will reach a Consensus

We found that the more cooperative strategies resulted in a higher *perceived* likelihood that the debaters could reach a consensus (β = 0.867, SE = 0.164, $p < 0.001$). The Compromising mediator (medium-cooperativeness, medium-assertiveness) was the most successful

Table 4: Pairwise comparison of *binary* topic opinion change (i.e., if post-test agreement differed from pre-test), across mediator strategies and control (Model A) and across the TKI *axis* (Model B)).

Contrast (across TKI *strategies* and *control*)	Estimate (log-odds)	Standard Error (SE)	p-value
Forceful - Accommodating	-0.443	0.284	0.627
Forceful - Averting	0.017	0.282	0.999
Forceful - Collaborating	-0.278	0.283	0.924
Forceful - Compromising	-0.362	0.287	0.807
Forceful - Control	0.593	0.288	0.309
Accommodating - Averting	0.460	0.287	0.596
Accommodating - Collaborating	0.165	0.287	0.993
Accommodating - Compromising	0.081	0.290	0.999
Accommodating - Control	**1.036**	**0.293**	**0.006**
Averting - Collaborating	-0.295	0.285	0.906
Averting - Compromising	-0.378	0.289	0.781
Averting - Control	0.576	0.291	0.353
Collaborating - Compromising	-0.083	0.290	0.999
Collaborating - Control	**0.871**	**0.293**	**0.035**
Compromising - Control	**0.955**	**0.296**	**0.016**
Contrast (across the TKI *axis*)	**Estimate (log-odds)**	**Standard Error (SE)**	**p-value**
High-cooperativeness - Low-cooperativeness	0.377	0.202	0.063
High-assertiveness - Low-assertiveness	0.048	0.202	0.814

for reducing the perception of audience polarisation compared to the Control 'no-mediator' group (β = 0.719, SE = 0.239, p < 0.05; when compared across all strategies in Table 5), followed by the Accommodating and the Collaborating mediators. The Forceful moderator had a perceived polarising effect *on the debaters* in addition to themselves (RQ2 - Figure 6), while the Averting moderator also performed poorly in perceived debater consensus-building even though it was relatively successful in depolarising the participants (RQ2 - Figure 6). We present all pairwise comparisons in Table 5 with the visualisations for each respective strategy shown in Figure 7.

The audience's personal TKI preference did not have a significant relationship in improving the impact of the mediator unlike the audience-targeted RQ1 PAS and RQ2 *personal* agreement. Other relevant predictors included that those with more experience in *writing/engaging in online debates* such as those who write 'sometimes over the past six months' compared to 'daily' were more likely to perceive that the debaters would reach a consensus (β = 0.805, SE = 0.376, p < 0.05). Conversely, those that frequently read online debates were less likely to perceive that the debaters would reach a consensus (i.e., reading online debates/discussions 'daily' compared to 'sometimes over the past six months'; β = 0.477, SE = 0.213, p < 0.05).

Furthermore, we aggregate the 7-pt Likert values into the trinomial categories of the belief that the debaters 'will not reach consensus' (Likert values 1-3), have an unsure 'Neutral view' (4) and the belief that the debaters 'will reach an agreement' (4-7). Figure 8 highlights that the mediator bots had a significant impact on the audience in reducing their perceived polarisation—thus helping achieve our aim to detoxify the polarised perception of social media. For the Control 'no-mediator' group, only 26.67% of participants

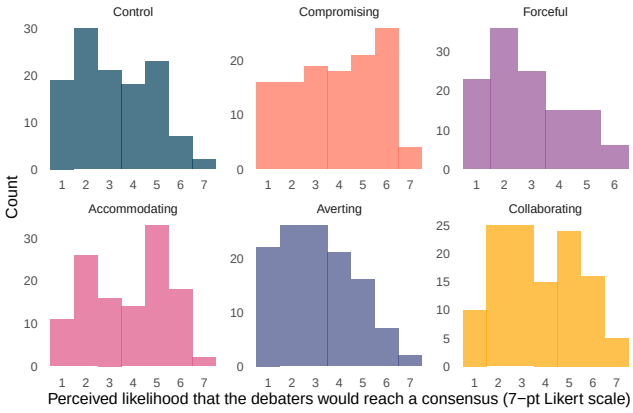

Figure 7: Distribution of the perceived likelihood that the debaters would reach a consensus (7-pt Consensus Likert) between the intervention and control groups.

believed that the debaters could reach a consensus compared to 44.17% for the Compromising mediator. Conversely, the Forceful mediator increased the perception of polarisation compared to the cooperative strategies and control—with only 17.5% believing that the debaters could reach a consensus.

4.3.1 *Co-opting Optimism vs. Nihilism.*

For the Compromising responses, users co-opted a belief that a compromise was necessary (even though the term never appeared in the transcripts). This was seen as being useful in cases of strong *perceived* disagreement between the debaters, such as perceiving

Table 5: Pairwise comparison of perceived likelihood that the debate will reach a consensus across mediator strategies and the control (Model A) and across the overall TKI axis (Model B).

Contrast (across TKI *strategies* and *control*)	Estimate	Standard Error (SE)	p-value
Forceful - Accommodating	**-1.030**	**0.231**	**<.001**
Forceful - Averting	-0.287	0.225	0.798
Forceful - Collaborating	**-1.008**	**0.231**	**<.001**
Forceful - Compromising	**-1.198**	**0.233**	**<.001**
Forceful - Control	-0.480	0.233	0.308
Accommodating - Averting	**0.743**	**0.229**	**0.015**
Accommodating - Collaborating	0.022	0.233	0.999
Accommodating - Compromising	-0.168	0.234	0.980
Accommodating - Control	0.551	0.235	0.178
Averting - Collaborating	**-0.721**	**0.229**	**0.021**
Averting - Compromising	**-0.911**	**0.232**	**0.001**
Averting - Control	-0.193	0.232	0.962
Collaborating - Compromising	-0.190	0.234	0.966
Collaborating - Control	0.529	0.237	0.224
Compromising - Control	**0.719**	**0.239**	**0.031**
Contrast (across the TKI *axis*)	**Estimate**	**Standard Error (SE)**	**p-value**
High-cooperativeness - Low-cooperativeness	**0.867**	**0.164**	**<.001**
High-assertiveness - Low-assertiveness	0.158	0.162	0.330

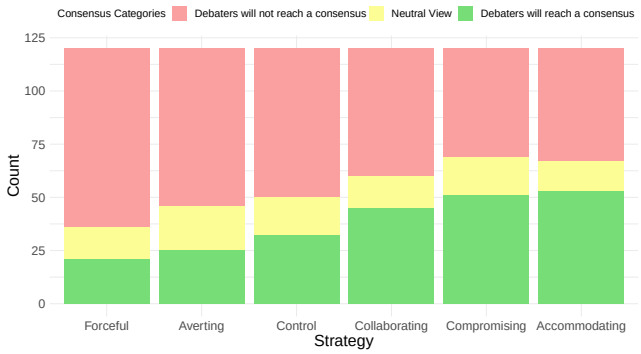

Figure 8: 7-pt Consensus Likert ratings in trinomial form as represented whether: the audience perceives that the debate will not reach a consensus (Likert values: 1-3), neither sure/unsure (Likert value: 4), and the debate will reach a consensus (Likert values: 5-7).

that *"The views are far apart but I think both realize there is no complete situation of either or. I think with concessions and compromise both could be satisfied"* (P29) and *"I think [Debaters] F and E, would eventually compromise on the issue as the mediator suggested. Both have valid points, and should take in account a solution that addresses both."* (P33).

Moreover, the high-cooperativeness mediators tended to lead to more *optimism* in the sentiment of the audience responses, such as stating that *"The mediator took the debate from only thinking about possible scenarios to reality, and I think that is very important. It's easy to create "what if's" in our minds with our own personal bias."* (P91). As well as seeing the mediator as a *"proactive approach to facilitating meaningful discussion and potential resolution"* (P107)—which stresses the aim to produce proactive interactive solutions to combat polarisation rather than reactive solutions from human moderators, such as banning, blocking, or bullying.

Conversely, our qualitative responses explain the Forceful mediator's strong impact on perceived polarisation (Figure 8) as the users co-opted the mediator's assertive and low-cooperative mentality. Specifically, the audience's responses when presented with the Forceful mediator resulted in them co-opting negative and destructive criticism, with participants claiming that the debaters' arguments were "very weak, and both can be easily dismissed" (P109) and that "there is no middle ground that they can inhabit" (P74).

5 DISCUSSION

In our work we aim to identify ways in which AI-powered bots can have a positive effect on the perceptions of discourse in online social media. The proliferation of chatbots and language models has raised fears that social media will soon be plagued by automated bots that spread rumours, misinformation, and lead to polarisation. Recent announcements regarding social bots by Meta [22] and Discord [5] have not curbed these concerns, and instead have fuelled further worries regarding privacy and addiction.

Our work has shown the potential benefits of psychologically-informed AI mediator-bots on users. Specifically, we choose to focus on debate audiences rather than debaters themselves, because rapid-fire online debates occur only briefly, whereas their transcripts exist in perpetuity for many to read. Hence, focusing on audiences ensures that the technology is relevant to more people, and is also 'backwards compatible' in the sense that our proposed bots can mediate debates that have already taken place.

The potential benefits are substantial. Primarily, we argue that the mediators we propose can address the harmful culture of social media—with its perceptions of insolvable divisions [91], toxic 'flame wars' [29, 78], and perceptions of nihilism due to a lack of cooperation and solutions [48, 73]. The voluntary implementation of AI mediators by social media platforms would reflect their commitments to a safe and open social media [12, 13, 24], where these mediators could be deployed as a client-side extension to users' browsers or as a voluntary feature in online threads.

Next, we outline the implications and applicability of our findings in regards to the mediators' persuasiveness (RQ1), psychological impact (RQ2), and for perceived division (RQ3). We then contextualise our findings in regards to our target practitioners of social media platforms (with regards to our prompt-tuning approach), alongside HCI and Psychology researchers. Finally, we reflect on potential future work to build on our findings and further contribute to improving social media culture via voluntary interventions.

5.1 Persuasiveness of Mediator Bots Strategies (RQ1)

While high-cooperativeness strategies were generally more persuasive, PAS was the only metric for which the assertiveness of the mediator had a moderate and significant effect due to the audience's perception that assertive behaviour imbues *confidence* (per our qualitative data). This corroborates with prior work on political (mis)information that showed that *confidence in information presentation* is a significant factor for improving vaccine uptake when using COVID-19 chatbots and conversational agents [41, 51], as well as for enhancing user trust of AI decisions [9, 10, 17, 41].

However, while moderate assertiveness is necessary to demonstrate *confidence*, high-assertive behaviour can detract the focus of the audience away from the debate itself and redirect it towards the mediator—as seen in the polarising/divisive effect of the Forceful mediator in Figure 7. The balance of assertiveness corroborates with riot control dynamics, whereby confrontational rhetoric can escalate tensions and provoke a defensive reaction from the protesting or rioting groups through creating a *majority vs authority* mob-mentality dynamic [46]. Similarly, in online discourse, the assertive behaviours from authoritative figures can trigger community-level radicalisation through a cascading *belief stampede* [25]. Thus, high-cooperativeness is an essential mediating force for assertiveness to avoid *polarising* and immolating the debate.

More generally, PAS can act as an overall 'gateway' variable for analysing argumentation, as it acts as a generalised argumentation and persuasion 'performance' measure. Critically, PAS aims at convincing and changing audience perceptions [4, 37, 81, 93]—which is required for the mediator's depolarising arguments to have its intended effect. We envision that PAS could become a standardised metric for future mediator-bot designs—where social media platforms or researchers could *substitute* the language model based on other argumentative NLP studies which utilise PAS as a measure for conversational language model performance [51, 81]. PAS is currently a common benchmark metric for analysing Public Service Announcement strategies [4, 38, 93]. Thus, PAS can be applied as an approximate benchmark for potential mediator-bot performance of a language model.

5.2 Mediator Strategy can Impact People's Beliefs (RQ2)

By design, content *moderation* must be *consistent*, *enforced*, and *reactive*—as the failure to uphold consistent flagging or removal of content can be perceived as a form of platform bias, which can lead to a sense of injustice and polarisation [28]. Content mediation intends to reduce perceptions of polarisation before they necessitate immediate removal, as well as reduce the toxic divides in modern discourse through the mediator's questioning and solution-building approaches.

Mediator-bots could act as proactive and preemptive tools to combat the degradation of online debates into polarised hostile groups which imbue current social media culture [11, 54]. Further, our mediator bots do not rely on stifling speech or controlling the discussion to reduce polarisation, unlike moderators—as evident by the impact of our more successful mediator-bots as shown in Figure 8, without any changes or even responses from the debaters. Thus, our argumentation-based approach to depolarisation adds to the toolkit for reducing online polarity by platforms, researchers, and the government.

Importantly, our approach underlines the commitment by platforms and governments to develop non-invasive interventions to combat polarisation and extremism agreed upon in the transnational Christchurch Call [12, 13]. As such, platforms could either implement mediator-bots across discussion threads—such as debate or news subreddits or utilise an external measure to justify *when* the mediator-bot should intervene. For instance, a mediator-bot could be contingent on keyword or sentiment-detection models to detect extreme polarity (such as extremism or hate-speech detection models [28, 29, 60]) and interject thereafter to utilise the calming presence experienced via the Accommodating strategy, or even for the Averting mediator to deescalate the debate and divert towards less controversial solutions.

For researchers, our findings extend on prior work by identifying that TKI preference is significant *and* also applies to external third parties (in this case mediators) in addition to the person themselves [76, 88]. Thus, platforms could utilise a one-off self-questionnaire on registration to *personalise* which mediator strategy they personally see to maximise the personal depolarisation effect. Moreover, researchers could extend our mediation strategy findings (e.g., Accommodating as the highest-performing strategy) to test *apolitical* mediation scenarios—such as testing the TKI-strategy mediator-bot for resolving personal online disputes or for persona-driven social bots (such as those proposed by Meta [22]), and for mental health support [78]. However, future work in human oversight of AI decision-making and discourse, as well as conversational ethics and its societal implications, is critical to ensure robust AI interventions that protect at-risk groups from AI mistakes or inadvertent risks such as alienation [34].

5.3 Perceived Consensus-Building in Online Debates (RQ3)

It is crucial to reduce the perception that debates are polarised and unsolvable, as it can negatively impact the audience's mental health through disengagement in social media discussions [19]. In addition, this has been shown to lead to dangerous existentialist

and nihilistic outlooks on society [48, 73]. In our work, we found a general trend that the cooperative strategies were more successful at increasing the perceived likelihood of debaters' consensus, with participants co-opting a more optimistic approach to consensus-building amongst debaters. This finding matches previous work that found that injecting more optimistic news into divisive conflicts moderately improved the self-reported optimism and positive experiences of social media users' [8]. Conversely, the Forceful mediators led participants to project more *negative* feelings, resulting in a reduction of participants' belief that the debaters could reach a consensus.

However, unlike the audience-targeted argument strength (PAS) and personal opinion, there was no significant relationship between preference for the TKI axis and any of the used strategy. The lack of TKI preference impact is likely due to the difference between *personal* opinion and *perceived (debaters)* opinions. Specifically, the reader's preferred TKI influences PAS values as PAS is an introspective judgement on the argument within the wider conflict, which relies on the reader's *personal view* on the mediator's points. Conversely, the lack of a link between preferred TKI strategy and perceived likelihood of the debaters reaching a consensus is likely a result of the participants' *projected view on how others would view the conflict-resolution strategy* rather than reflecting on their own view and preference.

5.4 AI Strategisation in Human-AI Collaboration

Our findings have important implications for social media users and platforms, and HCI/Psychology researchers. Firstly, our manipulation check highlights that Large Language Models (LLMs) can replicate specific psychological strategies with high confidence (with substantial inter-rater agreement). Researchers can benefit from these findings as they demonstrate that LLMs can be prompt-personalised to operationalise, automate, and test social and psychological theories at scale. Likewise, the findings that assertive and cooperative strategies improve argument strength could be useful for tools to improve formal *persuasive* writing—which could supplement or improve writing aides like Grammarly [30] through using psychologically-driven strategies. This prompt-driven approach can be utilised for creative ideation and dispute-resolution —such as utilising the Accommodating mediator as a bot in Slack or Teams workplace discussion, or in informal Reddit-like discussions. The hybridisation of formal psychology-driven *strategies* and an open-ended bot approach enables researchers and online communities to engage with AI as a collaborative partner in resolving online and hybrid disputes, ideation tasks, and persuasion needs—with the potential to extend the mediator-bot design as a future business or legal partner/tool.

Our findings advance mediation through testing conflict resolution strategies on a *third-party* rather than a debater—identifying that the audience have a slight preference for *a mediator which has a strategy that matches their preference* as well as the trend that individual grievance/disagreement solving (high-cooperative) strategies are more persuasive (RQ1), encourage critical thinking on pre-existing beliefs (RQ2), and more effective at increasing the plausibility for debater consensus (RQ3).

5.5 Limitations and Future Work

Our study seeks to introduce psychology-driven strategies for online debate depolarisation in a simulated environment with real-world data. However, our experimental design only considers the *audience* response rather than the potential *debaters* response. Our focus on just mediator replies without debater responses highlights the common scenario where a user skims a high-level subreddit topic thread rather than reading the individual in-depth reply chains. In context, prior work identified that reading online *argumentative text* reduced in-depth and focused reading compared to general expository text, whereby users tended to jump between top-level points/paragraphs in the argumentative text (i.e., "non-linear reading") [52].

Furthermore, the added benefit of this approach is that it avoids the confounding effect of positive or negative debater responses or tangents interfering with the mediator's strategies/content. By letting the audience evaluate the mediator's replies, we ensure that they judge the mediator's effectiveness in their depolarisation process based on their own criteria, not peer pressure and social conformity [85–87]. Nonetheless, future work should expand on our study by focusing on the role of the mediator in countering *emotive* debater responses—to test the robustness of the mediators in intra-debate disputes. Real-time experiments between debaters could also test the utility of conversational and context dynamics, as well as the mediator's ability to withstand and counter abusive responses or hate speech. In addition, future work in language model-based mediation targeting *debaters themselves* could involve exploring and enforcing specific moderator rules to improve consensus-building, such as the influential factors of conversation tempo [45], encouraging involvement of under-represented debaters [39, 87], and voting on the debate topic or mediator(s) responses [87].

The purpose of the mediator-bots for practitioners/platforms is not to replace but to supplement content-moderation systems. Our mediator-bots target the topic-*specific* ideas and concepts in a debate with the intent to promote critical thinking with our *Questioning*, *Claim* speech acts; and promote consensus through our *Respond* speech acts (such as the proposed solution shown in our example Figure 2). We envision future work to highlight *trigger-mechanisms* to detect when to engage in content-*moderation* to regulate the tone, semantics, and conduct of a debate as a second line of defence when the mediator-bots *idea*-driven *depolarisation* techniques must be supplemented with *emotive*-driven *deescalation* techniques. However, content-mediation should not stymie or override debate as polarised discussions are not always negative. Thus, mediation suits voluntary opt-in collaborative *idea-building and critical thinking* while moderation is suited for *regulating and deescalating* speech that has strayed from the platform or subreddit/thread community's rules (such as addressing trolling or countering hate speech).

Future work should also consider interdisciplinary evaluations of whether AI mediation should pursue *neutrality* vs. *principles*—a topic which remains unresolved even in the field of mediation itself [74]. Moreover, the presentation of information (facts and sources pulled by GPT-4 [57]), as well as the perspective of the mediator, may not necessitate 'a compromise' or even a debate. In these cases, this is where human oversight and content-*moderation* comes into play—as a debate that violates rights/human dignity

(e.g., violent/hate speech) likely mandates a *reactive* moderation rather than *proactive* mediation solution. Thus, future work should consider the potential ethics of language models and their own cognitive/information biases in AI collaboration activities—and create prompt-engineering methods to counter-bias LLM responses.

6 CONCLUSION

Content mediation offers an additional collaborative avenue to depolarise social media audiences on divisive topics. We investigate a prompt-tuned GPT-4 mediation design based on the five TKI conflict resolution psychological strategies. Our findings highlight the effectiveness of different strategies in terms of their Perceived Argument Strength (PAS), their depolarising effect on the audience of a debate, and the audience's perceived likelihood that the debaters will come together and reach a consensus. We present our designs, prompts and analysis framework to enable social media platforms to consider new *proactive* depolarisation measures before resorting to *reactive* moderation, which can be perceived as heavy-handed and stifling. Our findings also highlight that user preferences towards the TKI cooperativeness/assertiveness axis improves the mediator's argument persuasiveness and depolarising effect, enabling platforms to consider stylising mediator-bots to optimise and enhance users' interaction in online discussions. Our results indicate that mediator-bots can improve the user experience of social media through detoxifying the hostile and polarised culture of online discourse. By bridging the divide of modern social media discourse, we can mitigate the negative mental health effects caused by online polarisation, all while protecting open and secure speech.

REFERENCES

[1] Francis M. Adams and Charles E. Osgood. 1973. A Cross-Cultural Study of the Affective Meanings of Color. *Journal of Cross-Cultural Psychology* 4, 2 (1973), 135–156. https://doi.org/10.1177/002202217300400201

[2] Reginald Adkins. 2015. *Conflict Management Styles Quiz.* North Carolina State University. https://facultyombuds.ncsu.edu/files/2015/11/Conflict-management-styles-quiz.pdf

[3] Jungyong Ahn, Jungwon Kim, and Yongjun Sung. 2022. The effect of gender stereotypes on artificial intelligence recommendations. *Journal of Business Research* 141 (2022), 50–59. https://doi.org/10.1016/j.jbusres.2021.12.007

[4] Elisabeth Bigsby, Joseph N. Cappella, and Holli H. Seitz. 2013. Efficiently and Effectively Evaluating Public Service Announcements: Additional Evidence for the Utility of Perceived Effectiveness. *Communication Monographs* 80, 1 (2013), 1–23. https://doi.org/10.1080/03637751.2012.739706

[5] Matt Binder. 2023. *Discord is rolling out new features powered by AI.* Discord. https://mashable.com/article/discord-ai-features-announcements/

[6] Robert R. Blake, Jane S. Mouton, and Alvin C. Bidwell. 1962. Managerial grid. *Advanced Management - Office Executive* 1, 9 (1962), 12–15.

[7] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[8] Kathryn Buchanan, Lara B. Aknin, Shaaba Lotun, and Gillian M. Sandstrom. 2021. Brief exposure to social media during the COVID-19 pandemic: Doom-scrolling has negative emotional consequences, but kindness-scrolling does not. *PLOS ONE* 16, 10 (10 2021), 1–12. https://doi.org/10.1371/journal.pone.0257728

[9] Francisco Maria Calisto, João Fernandes, Margarida Morais, Carlos Santiago, João Maria Abrantes, Nuno Nunes, and Jacinto C. Nascimento. 2023. Assertiveness-Based Agent Communication for a Personalized Medicine on Medical Imaging Diagnosis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 13, 20 pages. https://doi.org/10.1145/3544548.3580682

[10] C. Castelfranchi and R. Falcone. 1998. Principles of Trust for MAS: Cognitive Anatomy, Social Importance, and Quantification. In *Proceedings of the 3rd International Conference on Multi Agent Systems (ICMAS '98)*. IEEE Computer Society, USA, 72.

[11] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. *The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data.* Association for Computing Machinery, New York, NY, USA, 3175––3187. https://doi.org/10.1145/3025453.3026018

[12] Christchurch Call. 2021. Algorithms & Positive Interventions Workplan 2021. https://www.christchurchcall.com/assets/Documents/Algorithms-and-Positive-Interventions-WorkPlan.pdf

[13] Christchurch Call. 2021. The Christchurch Call to Action To Eliminate Terrorist and Violent Extremist Content Online. https://www.christchurchcall.com/assets/Documents/Christchurch-Call-full-text-English.pdf

[14] Jacob Cohen. 1992. Statistical Power Analysis. *Current Directions in Psychological Science* 1, 3 (1992), 98–101. http://www.jstor.org/stable/20182143

[15] L.M. Collins. 2007. Research Design and Methods. In *Encyclopedia of Gerontology (Second Edition)* (second edition ed.), James E. Birren (Ed.). Elsevier, New York, NY, 433–442. https://doi.org/10.1016/B0-12-370870-2/00162-1

[16] Milo Comerford, Jakob Guhl, and Jack Miller. 2021. *Understanding the New Zealand Online Extremist Ecosystem.* Institute for Strategic Dialogue, London, UK.

[17] Samuel Rhys Cox and Wei Tsang Ooi. 2022. Does Chatbot Language Formality Affect Users' Self-Disclosure?. In *Proceedings of the 4th Conference on Conversational User Interfaces* (Glasgow, United Kingdom) *(CUI '22)*. Association for Computing Machinery, New York, NY, USA, Article 1, 13 pages. https://doi.org/10.1145/3543829.3543831

[18] Sauvik Das and Adam Kramer. 2021. Self-Censorship on Facebook. *Proceedings of the International AAAI Conference on Web and Social Media* 7, 1 (Aug. 2021), 120–127. https://doi.org/10.1609/icwsm.v7i1.14412

[19] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2021. Predicting Depression via Social Media. *Proceedings of the International AAAI Conference on Web and Social Media* 7, 1 (Aug. 2021), 128–137. https://doi.org/10.1609/icwsm.v7i1.14432

[20] George Denison. 2023. *How much should you pay research participants?* Prolific. Retrieved May 14, 2023 from https://www.prolific.co/blog/how-much-should-you-pay-research-participants

[21] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, MI, 2368–2378. https://doi.org/10.18653/v1/N19-1246

[22] Benji Edwards. 2023. *Meta plans AI-powered chatbots to boost social media numbers.* Ars Technica. https://arstechnica.com/information-technology/2023/08/meta-readies-ai-chatbots-for-artificial-companionship-and-user-retention/

[23] Facebook. 2023. *Facebook Community Standards.* Meta. Retrieved September 4, 2023 from https://transparency.fb.com/en-gb/policies/community-standards/

[24] Facebook. 2023. *Our Commitment to Safety.* Meta. Retrieved September 4, 2023 from https://www.facebook.com/business/news/our-commitment-to-safety

[25] Philip Feldman. 2023. Chapter Fifteen - Belief stampedes. In *Stampede Theory*, Philip Feldman (Ed.). Elsevier, Cambridge, MA, 217–219. https://doi.org/10.1016/B978-0-44-313735-8.00024-3

[26] Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76, 5 (1971), 378–382. https://doi.org/10.1037/h0031619

[27] Yoshiko Goda, Masanori Yamada, Hideya Matsukawa, Kojiro Hata, and Seisuke Yasunami. 2014. Conversation with a Chatbot before an Online EFL Group Discussion and the Effects on Critical Thinking. *The Journal of Information and Systems in Education* 13 (12 2014), 1–7. https://doi.org/10.12937/ejsise.13.1

[28] Jarod Govers, Philip Feldman, Aaron Dant, and Panos Patros. 2023. Down the Rabbit Hole: Detecting Online Extremism, Radicalisation, and Politicised Hate Speech. *ACM Comput. Surv.* 55, 14s, Article 319 (jul 2023), 35 pages. https://doi.org/10.1145/3583067

[29] Jarod Govers, Philip Feldman, Aaron Dant, and Panos Patros. 2023. PromptGAN–Customisable Hate Speech and Extremist Datasets via Radicalised Neural Language Models. In *Proceedings of the 2023 9th International Conference on Computing and Artificial Intelligence* (Tianjin, China) *(ICCAI '23)*. Association for Computing Machinery, New York, NY, USA, 515–522. https://doi.org/10.1145/3594315.3594366

[30] Grammarly. 2023. *AI Writing Assistance.* Retrieved November 22, 2023 from https://www.grammarly.com/ai

[31] Ted Grover and Gloria Mark. 2019. Detecting Potential Warning Behaviors of Ideological Radicalization in an Alt-Right Subreddit. *Proceedings of the International AAAI Conference on Web and Social Media* 13, 01 (Jul. 2019), 193–204. https://ojs.aaai.org/index.php/ICWSM/article/view/3221

[32] Zixuan Guo and Tomoo Inoue. 2019. Using a Conversational Agent to Facilitate Non-Native Speaker's Active Participation in Conversation. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3313075

[33] Jay Hall. 1969. *Conflict management survey: A survey of one's characteristic reaction to and handling of conflicts between himself and others.* Teleometrics International., The Woodlands, TX.

[34] Jeffrey T Hancock, Mor Naaman, and Karen Levy. 2020. AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. *Journal of Computer-Mediated Communication* 25, 1 (01 2020), 89–100. https://doi.org/10.1093/jcmc/zmz022

[35] Jess Hohenstein and Malte Jung. 2020. AI as a Moral Crumple Zone: The Effects of AI-Mediated Communication on Attribution and Trust. *Comput. Hum. Behav.* 106, C (may 2020), 13 pages. https://doi.org/10.1016/j.chb.2019.106190

[36] George Caspar Homans. 1961. *Social behavior: Its elementary forms.* Harcourt, Brace, Oxford, England. 404–404 pages.

[37] Jos Hornikx, Annemarie Weerman, and Hans Hoeken. 2022. An exploratory test of an intuitive evaluation method of perceived argument strength. *Studies in Communication Sciences* 22, 2 (2022), 311–324. https://doi.org/10.24434/j.scoms.2022.02.003

[38] Irina Alexandra Iles and Xiaoli Nan. 2017. It's no laughing matter: An exploratory study of the use of ironic versus sarcastic humor in health-related advertising messages. *Health Marketing Quarterly* 34, 3 (2017), 187–201. https://doi.org/10.1080/07359683.2017.1346432

[39] Hyo Jin Do, Ha-Kyung Kong, Pooja Tetali, Karrie Karahalios, and Brian P. Bailey. 2023. Inform, Explain, or Control: Techniques to Adjust End-User Performance Expectations for a Conversational Agent Facilitating Group Chat Discussions. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 343 (Oct. 2023), 26 pages.

[40] Jillian Christine Johnson. 2022. Paranoid Posting: An Analysis of Being Too Online. *SIGCAS Comput. Soc.* 50, 2 (aug 2022), 16–17. https://doi.org/10.1145/3557805.3557814

[41] Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T. Hancock. 2023. Working With AI to Persuade: Examining a Large Language Model's Ability to Generate Pro-Vaccination Messages. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 116 (apr 2023), 29 pages. https://doi.org/10.1145/3579592

[42] Rune Karlsen, Kari Steen-Johnsen, Dag Wollebæk, and Bernard Enjolras. 2017. Echo chamber and trench warfare dynamics in online debates. *European Journal of Communication* 32, 3 (2017), 257–273. https://doi.org/10.1177/0267323117695734

[43] Kialo. 2023. *Kialo - Explore Debates.* Retrieved Aug 29, 2023 from https://www.kialo.com/

[44] Ralph H. Kilmann and Kenneth W. Thomas. 1977. Developing a Forced-Choice Measure of Conflict-Handling Behavior: The "Mode" Instrument. *Educational and Psychological Measurement* 37, 2 (1977), 309–325. https://doi.org/10.1177/001316447703700204

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

[46] Graig R. Klein and Patrick M. Regan. 2018. Dynamics of Political Protests. *International Organization* 72, 2 (2018), 485–521. https://www.jstor.org/stable/26569481

[47] Richard A. Klein. 2018. Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances in Methods and Practices in Psychological Science* 1, 4 (2018), 443–490. https://doi.org/10.1177/2515245918810225

[48] K. H. Kristinsdottir, H. F. Gylfason, and R. Sigurvinsdottir. 2021. Narcissism and Social Media: The Role of Communal Narcissism. *Int J Environ Res Public Health* 18, 19 (2021), 1660–4601. https://doi.org/10.3390/ijerph181910106

[49] P.R. Lawrence and J.W. Lorsch. 1967. *Organization and Environment: Managing Differentiation and Integration.* Division of Research, Graduate School of Business Administration, Harvard University, Cambridge, MA. https://books.google.com.au/books?id=VlJIAAAAMAAJ

[50] Jae Kook Lee, Jihyang Choi, Cheonsoo Kim, and Yonghwan Kim. 2014. Social Media, Network Heterogeneity, and Opinion Polarization. *Journal of Communication* 64, 4 (01 2014), 702–722. https://doi.org/10.1111/jcom.12077

[51] Kristi Yoonsup Lee, Saudamini Vishwanath Dabak, Vivian Hanxiao Kong, Minah Park, Shirley L. L. Kwok, Madison Silzle, Chayapat Rachatan, Alex Cook, Aly Passanante, Ed Pertwee, Zhengdong Wu, Javier A. Elkin, Heidi J. Larson, Eric H. Y. Lau, Kathy Leung, Joseph T. Wu, and Leesa Lin. 2023. Effectiveness of chatbots on COVID vaccine confidence and acceptance in Thailand, Hong Kong, and Singapore. *npj Digital Medicine* 6, 1 (2023), 96. https://doi.org/10.1038/s41746-023-00843-6

[52] Ziming Liu. 2005. Reading behavior in the digital environment. *Journal of Documentation* 61, 6 (2005), 700–712. https://doi.org/10.1108/00220410510632040

[53] Andrey Lovakov and Elena R. Agadullina. 2021. Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology* 51, 3 (2021), 485–504. https://doi.org/10.1002/ejsp.2752

[54] Dillon Ludemann. 2018. /pol/emics: Ambiguity, scales, and digital discourse on 4chan. *Discourse, Context & Media* 24 (2018), 92–98. https://doi.org/10.1016/j.dcm.2018.01.010

[55] Alice E. Marwick and danah boyd. 2011. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society* 13, 1 (2011), 114–133. https://doi.org/10.1177/1461444810365313 arXiv:https://doi.org/10.1177/1461444810365313

[56] Teale W. Masrani, Jack Jamieson, Naomi Yamashita, and Helen Ai He. 2023. Slowing It Down: Towards Facilitating Interpersonal Mindfulness in Online Polarizing Conversations Over Social Media. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 90 (apr 2023), 27 pages. https://doi.org/10.1145/3579523

[57] Yusuf Mehdi. 2023. *Confirmed: the new Bing runs on OpenAI's GPT-4.* Microsoft. Retrieved Aug 29, 2023 from https://blogs.bing.com/search/march_2023/Confirmed-the-new-Bing-runs-on-OpenAI%E2%80%99s-GPT-4

[58] David G. Myers and Helmut Lamm. 1976. The group polarization phenomenon. *Psychological Bulletin* 83, 4 (1976), 602–627. https://doi.org/10.1037/0033-2909.83.4.602

[59] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[60] Sebastião Pais, Irfan Khan Tanoli, Miguel Albardeiro, and João Cordeiro. 2020. Unsupervised Approach to Detect Extreme Sentiments on Social Networks. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).* IEEE Press, The Hague, Netherlands, 651–658. https://doi.org/10.1109/ASONAM49781.2020.9381420

[61] Henry Prakken. 2000. On Dialogue Systems with Speech Acts, Arguments, and Counterarguments. In *Proceedings of the European Workshop on Logics in Artificial Intelligence (JELIA '00).* Springer-Verlag, Berlin, Heidelberg, 224–238.

[62] ProCon. 2023. *ProCon.org - Pros and Cons of 100+ Topics.* Britannica Group. Retrieved Aug 29, 2023 from https://www.procon.org/

[63] Geetanjali Rakshit, Kevin K. Bowden, Lena Reed, Amita Misra, and Marilyn Walker. 2019. *Debbie, the Debate Bot of the Future.* Springer International Publishing, Cham, 45–52. https://doi.org/10.1007/978-3-319-92108-2_5

[64] John O. Rawlings, Sastry G. Pantula, and David A. Dickey (Eds.). 1998. *Class Variables in Regression.* Springer New York, New York, NY, 269–323. https://doi.org/10.1007/0-387-22753-9_9

[65] Richard Rogers. 2020. Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media. *European Journal of Communication* 35, 3 (2020), 213–229. https://doi.org/10.1177/0267323120922066

[66] Kavous Salehzadeh Niksirat, Diana Korka, Hamza Harkous, Kévin Huguenin, and Mauro Cherubini. 2023. On the Potential of Mediation Chatbots for Mitigating Multiparty Privacy Conflicts - A Wizard-of-Oz Study. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 142 (apr 2023), 33 pages. https://doi.org/10.1145/3579618

[67] Andrea Schneider and Jennifer Brown. 2013. Negotiation Barometry: A Dynamic Measure of Conflict Management Style. *Ohio State University Journal On Dispute Resolution* 11 (04 2013).

[68] Thomas Schäfer and Marcus A. Schwarz. 2019. The Meaningfulness of Effect Sizes in Psychological Research: Differences Between Sub-Disciplines and the Impact of Potential Biases. *Frontiers in Psychology* 10 (2019), 13 pages. https://doi.org/10.3389/fpsyg.2019.00813

[69] Joseph Seering. 2020. Reconsidering Self-Moderation: The Role of Research in Supporting Community-Based Models for Online Content Moderation. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 107 (oct 2020), 28 pages. https://doi.org/10.1145/3415178

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

[71] Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, et al. 2021. An autonomous debating system. *Nature* 591, 7850 (2021), 379–384.

[72] Min-Hsin Su, Jiyoun Suk, and Hernando Rojas. 2022. Social Media Expression, Political Extremity, and Reduced Network Interaction: An Imagined Audience Approach. *Social Media + Society* 8, 1 (2022), 20563051211069056. https://doi.org/10.1177/20563051211069056 arXiv:https://doi.org/10.1177/20563051211069056

[73] Chris Sumner, Alison Byers, Rachel Boochever, and Gregory J. Park. 2012. Predicting Dark Triad Personality Traits from Twitter Usage and a Linguistic Analysis of Tweets. In *Proceedings of the 2012 11th International Conference on Machine Learning and Applications - Volume 02 (ICMLA '12).* IEEE Computer Society, USA, 386–393. https://doi.org/10.1109/ICMLA.2012.218

[74] Lawrence Susskind, Sarah McKearnan, and Jennifer Thomas-Larmer. 1999. The Consensus Building Handbook: A Comprehensive Guide to Reaching Agreement. https://doi.org/10.4135/9781452231389

[75] Kenneth W. Thomas and Ralph H. Kilmann. 1976. Thomas-Kilmann Conflict Mode Instrument. *Group & Organization Studies* 1, 2 (1976), 249–251. https://doi.org/10.1177/105960117600100214

[76] Kenneth W. Thomas and Ralph H. Kilmann. 1978. Comparison of Four Instruments Measuring Conflict Behavior. *Psychological Reports* 42, 3_suppl (1978), 1139–1145. https://doi.org/10.2466/pr0.1978.42.3c.1139

[77] Carlos Toxtli, Andrés Monroy-Hernández, and Justin Cranshaw. 2018. Understanding Chatbot-Mediated Task Management. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18).* Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3173574.3173632

[78] Filippo Trevisan. 2020. "Do You Want to Be a Well-Informed Citizen, or Do You Want to Be Sane?" Social Media, Disability, Mental Health, and Political Marginality. *Social Media + Society* 6, 1 (2020), 2056305120913909. https://doi.org/10.1177/2056305120913909

[79] John C. Turner and Penelope J. Oakes. 1986. The significance of the social identity concept for social psychology with reference to individualism, interactionism and social influence. *British Journal of Social Psychology* 25, 3 (1986), 237–252. https://doi.org/10.1111/j.2044-8309.1986.tb00732.x

[80] Greg Walsh and Eric Wronsky. 2019. AI + Co-Design: Developing a Novel Computer-Supported Approach to Inclusive Design. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing* (Austin, TX, USA) *(CSCW '19)*. Association for Computing Machinery, New York, NY, USA, 408–412. https://doi.org/10.1145/3311957.3359456

[81] Thiemo Wambsganss, Tobias Kueng, Matthias Soellner, and Jan Marco Leimeister. 2021. ArgueTutor: An Adaptive Dialog-Based Learning System for Argumentation Skills. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 683, 13 pages. https://doi.org/10.1145/3411764.3445781

[82] Klaus Weber, Niklas Rach, Wolfgang Minker, and Elisabeth André. 2020. How to Win Arguments. *Datenbank-Spektrum* 20, 2 (2020), 161–169. https://doi.org/10.1007/s13222-020-00345-9

[83] Senuri Wijenayake, Jolan Hu, Vassilis Kostakos, and Jorge Goncalves. 2021. Quantifying the Effects of Age-Related Stereotypes on Online Social Conformity. In *Human-Computer Interaction – INTERACT 2021*, Carmelo Ardito, Rosa Lanzilotti, Alessio Malizia, Helen Petrie, Antonio Piccinno, Giuseppe Desolda, and Kori Inkpen (Eds.). Springer International Publishing, Cham, 451–475.

[84] Senuri Wijenayake, Niels van Berkel, Vassilis Kostakos, and Jorge Goncalves. 2019. Measuring the Effects of Gender on Online Social Conformity. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 145 (nov 2019), 24 pages. https://doi.org/10.1145/3359247

[85] Senuri Wijenayake, Niels van Berkel, Vassilis Kostakos, and Jorge Goncalves. 2020. Impact of contextual and personal determinants on online social conformity. *Computers in Human Behavior* 108 (2020), 106302. https://doi.org/10.1016/j.chb.2020.106302

[86] Senuri Wijenayake, Niels van Berkel, Vassilis Kostakos, and Jorge Goncalves. 2020. Quantifying the Effect of Social Presence on Online Social Conformity. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 55 (may 2020), 22 pages. https://doi.org/10.1145/3392863

[87] Senuri Wijenayake, Niels van Berkel, Vassilis Kostakos, and Jorge Goncalves. 2021. Quantifying Determinants of Social Conformity in an Online Debating Website. *International Journal of Human-Computer Studies* 158 (11 2021), 102743. https://doi.org/10.1016/j.ijhcs.2021.102743

[88] Deanna F. Womack. 1988. Assessing the Thomas-Kilmann Conflict Mode Survey. *Management Communication Quarterly* 1, 3 (1988), 321–349. https://doi.org/10.1177/0893318988001003004

[89] Riska Wulandari, Baedhowi, and Aniek Hindrayani. 2021. Measuring Critical Thinking Skills with the RED Model. *Journal of Physics: Conference Series* 1808, 1 (mar 2021), 012030. https://doi.org/10.1088/1742-6596/1808/1/012030

[90] Davies Wyn and Matt Stevens. 2019. *The Importance of Critical Thinking and How to Measure It.* Pearson. Retrieved Aug 29, 2023 from https://www.talentlens.com/content/dam/school/global/Global-Talentlens/uk/AboutUs/Whitepapers/The-Importance-of-Critical-Thinking-and-How-to-Measure-It_UK_Final.pdf

[91] Moran Yarchi, Christian Baden, and Neta Kligler-Vilenchik. 2021. Political Polarization on the Digital Sphere: A Cross-platform, Over-time Analysis of Interactional, Positional, and Affective Polarization on Social Media. *Political Communication* 38, 1-2 (2021), 98–139. https://doi.org/10.1080/10584609.2020.1785067

[92] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a Machine Really Finish Your Sentence?. Retrieved from: https://aclanthology.org/P19-1472. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 4791–4800. https://doi.org/10.18653/v1/P19-1472

[93] X. Zhao, A. Strasser, J. N. Cappella, C. Lerman, and M. Fishbein. 2011. A Measure of Perceived Argument Strength: Reliability and Validity. *Commun Methods Meas* 5, 1 (2011), 48–75. https://doi.org/10.1080/19312458.2010.547822