Narratives and Perspectives: How AI Summaries Steer Users' Opinions and Engagement on Social Media

Jarod Govers
School of Computing and Information Systems
University of Melbourne
Melbourne, Victoria, Australia
jarod.govers@unimelb.edu.au

Cherie Sew
School of Computing and Information Systems
University of Melbourne
Melbourne, Victoria, Australia
csew@student.unimelb.edu.au

Eduardo Velloso
School of Computer Science
University of Sydney
Sydney, New South Wales, Australia
eduardo.velloso@sydney.edu.au

Vassilis Kostakos
School of Computing and Information Systems
University of Melbourne
Melbourne, Victoria, Australia
vassilis.kostakos@unimelb.edu.au

Jorge Goncalves
School of Computing and Information Systems
University of Melbourne
Melbourne, Victoria, Australia
jorge.goncalves@unimelb.edu.au

Abstract

AI summaries on social media are reshaping how users form opinions about political topics, yet their influence remains largely unexamined despite their widespread deployment. This paper investigates how two types of AI summaries affect user opinions and engagement: textual summaries of discussion narratives and percentage breakdowns of agreement/disagreement. Through a 144-participant experiment on simulated online discussion threads, we found that displaying commenter agreement percentages amplified social conformity towards the majority views beyond reading comments alone. Conversely, AI narrative summaries created misperceptions of balance in polarised threads, reducing opinion change. While these summaries did not influence participants' willingness to engage, toxic discussions deterred participation even when participants held majority views. Based on our findings, we provide critical design interventions for industry and researchers to mitigate these tools' polarising effects, paving the way for responsible AI deployment on social media platforms.

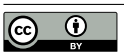
CCS Concepts

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

Keywords

Generative AI, social media, polarisation, online discourse, AI summary

ACM Reference Format:
Jarod Govers, Cherie Sew, Eduardo Velloso, Vassilis Kostakos, and Jorge Goncalves. 2026. Narratives and Perspectives: How AI Summaries Steer Users' Opinions and Engagement on Social Media. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3772318.3790945



CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3790945

1 Introduction

Generative Artificial Intelligence is now ubiquitous online, with social media platforms integrating it into nearly every aspect of social interaction. From helping users sift for news information [27, 58, 67, 68], to generating personalised insights into online information and discussions [60, 89], and filtering millions of posts and threads for relevance [114]. Thus, these AI tools are increasingly shaping the way people consume and engage with information online. Its capacity to distil large volumes of data into concise, digestible summaries has the potential to help detoxify social discourse and reduce the burden of engaging with hostile or overwhelming content. For example, when attempting to gauge public sentiment on a polarising issue, should a user be expected to wade through pages of vitriolic comments, or could an AI-generated summary serve as a more constructive and accessible alternative?

Tools that attempt to address this by displaying political alignment and presenting multiple viewpoints are not new. Platforms like Ground News already help users recognise bias and echo chambers by offering side-by-side perspectives on news stories [31]. Such approaches, when thoughtfully applied, can promote a better understanding of public sentiment, improve media literacy, and ultimately support a more informed and democratic public discourse [10, 27, 28, 107]. However, our growing reliance on AI to 'simplify' the information overload of social media places control in the hands of a small group of increasingly politicised companies [27]. This raises concerns about the potential for these systems to subtly (or overtly) influence public opinion and distort users' interpretations of posts and comments.

As platforms like Meta and Reddit integrate AI-generated summaries into comment sections, it becomes increasingly important to examine how these tools might influence public opinion. Moreover, evaluating users' ability to recognise and escape ideological echo chambers, with or without the assistance of AI, offers insight into the potential risks of AI-driven ideological steering. While concerns

about AI's influence on beliefs were explored in the context of interactive chatbots [27, 28, 58], there remains a significant gap in our understanding of how non-interactive AI tools, such as automated comment summaries, may contribute to subtle forms of ideological indoctrination on social media platforms.

Hence, we investigate whether AI-produced summaries, including textual summaries of viewpoints (henceforth AI *narratives*) and displaying the percentages of agreement on a thread's topic (henceforth AI *percentages*), can shift the polarity of user opinions in both civil and toxic discussions. Specifically, we examine whether such summaries (AI *narratives* and/or *percentages*) can subtly shift users' stances on controversial issues, depending on how perspectives and consensus are framed. In addition, we investigate how these AI summaries affect users' intention to engage with the discussion, across both civil and toxic discourse. Since AI-generated summaries may foreground certain arguments or highlight dominant viewpoints, they could shape the talking points users consider worth responding to, thereby altering not only what users think, but how and whether they choose to participate in a discussion at all.

In summary, our work aims to address the following Research Questions (RQs):

- RQ1 How do AI-generated summaries of ideological perspectives in social media threads influence users' opinions?
- RQ2 To what extent do AI-generated summaries of ideological perspectives impact users' likelihood of engaging with a social media thread?

We conducted an online experiment with 144 participants in which they were assigned to one of four conditions: a group exposed to both *AI narratives and AI percentages*, an *AI narratives only* group, an *AI percentages only* group, and a control without any AI summary tools. Participants were asked to read six simulated Reddit threads on a variety of topics, three with civil discourse and three with toxic discourse. Across each thread, participants saw different proportions of comments that agreed or disagreed with the overall post, and (outside of the control condition) saw their AI narratives' textual summary and/or the percentages of agreement/disagreement (Figure 1).

Our findings show that both AI tools significantly influenced participants' opinion change (RQ1), but their effects varied by type. Displaying AI-generated percentages of commenter agreement drove conformity, increasing the magnitude of opinion change toward the majority view when the comment section held polarised views compared to the control comments-only thread. Conversely, AI narrative summaries alone had a moderating effect; by presenting opposing arguments, they led participants to inaccurately perceive polarised debates as balanced, which in turn reduced their opinion shift. In addition, the combination of both tools led to the largest magnitude of opinion change within our participants.

The AI tools had no significant effect on participants' willingness to engage with the discussions (RQ2). Instead, we identified that the toxic social media threads (Department of Government Efficiency (DOGE), gun control, and the Ukraine war, which all included toxic discourse) were a powerful deterrent. Participants were more willing to engage in civil conversations, even when their stance was in the minority, versus toxic ones, even when their stance was in the majority. This highlights that constructive civil discourse is a more fundamental driver of participation than the perception of consensus.

Overall, we empirically demonstrate that AI summary tools can influence user behaviour by reinforcing consensus more than just reading the comments alone. Based on our findings, we outline design considerations for leveraging AI tools to facilitate detoxification, as well as strategies to limit the influence of AI summary tools. We highlight how responsible AI starts with identifying the potential systemic harms brought about by new AI tools *before* they are deployed. This serves as a call for action for industry, researchers, and regulators to consider how passive AI tools can exacerbate online polarisation.

2 Related Work

To contextualise the significance of opinion polarity and one's willingness to engage in online social media threads, we outline the psychological effects of social and informational conformity, alongside the current approaches of Large Language Model (LLM)-driven summaries on social media.

2.1 Psychological Dynamics of Online Public Opinion

Online discourse, especially on social media, plays a significant role in shaping public opinion and decision-making. There are three central psychological mechanisms that have been shown to underpin this influence: *conformity* (both social and informational), *pluralistic ignorance*, and *the spiral of silence theory*.

Asch's foundational work on conformity showed that individuals tend to align with majority views either to fit in (normative conformity) or because they perceive others as better informed (informational conformity) [4]. However, conformity research in social media typically focuses on topics relating to mis/disinformation. For example, Wijenayake et al. [110] identified that users who read a fake news story with comments critical of the content were more likely to have a negative attitude towards the fake news story, and were more likely to downvote it. Moreover, they identified that users were more likely to engage and reshare misinformation posts where everyone supported the premise of the post. However, this study only considered instances where comments either all support or all oppose the main post. Furthermore, Govers et al. [28] found that AI mediator bots can induce informational conformity by fostering the perception of consensus in polarised debates.

Unfortunately, users often misjudge the distribution of opinions online. Prior HCI studies have identified this cognitive bias, known as the *false consensus effect*, where users struggle to identify the majority opinion in threads where they themselves hold strong preconceptions—such as for support for political beliefs as the 'silent majority' [56, 96]. Conversely, users can also underestimate the popularity of their beliefs via *pluralistic ignorance*, where individuals wrongly believe their private views are in the minority, particularly if they believe that their beliefs are being persecuted/targeted by the online audience [87]. Moreover, echo chambers amplify these misconceptions [25, 41], especially on contentious topics like climate change [51, 61], populism [11, 115], and news bias [15, 27].

Another factor in opinion expression is the *spiral of silence theory*: individuals may withhold views if they perceive them to be unpopular [64]. For example, Hampton et al. [32] identified that only 42% of Facebook and 𝕏 (formerly Twitter) users were comfortable posting about a Snowden–NSA surveillance story online compared to 86% of Americans being willing to have such a conversation offline. Furthermore, Dubois and Szwarc [20] identified that users across France, Germany, UK and the US self-censored based on their own *perception* of whether they believed their views were in the minority.

These challenges are compounded by toxic or emotionally charged discourse, particularly on polarising topics. AI models, though not immune to bias [62], may help mitigate the effects of these psychological mechanisms by offering emotionally neutral summaries of discussions. It remains unclear, however, whether AI-generated summaries that display the percentage/proportions of agreement and disagreement on a social media thread would influence users' judgements or improve their awareness of the thread's polarisation. This is increasingly relevant as AI narrative summaries of news/discussions become embedded in social platforms [60, 68], news media [27, 58], and search engines [92].

In this context, Mun et al. [63] proposed AI-driven counterspeech and summarisation tools, cautioning that reactive AI engagement may intensify polarisation. Instead, they advocate for reflective summarisation through neutral, context-aware overviews that present diverse perspectives without reinforcing division. Our study builds on this work by assessing the influence of AI-generated summaries, both of narrative content and commenter sentiment distribution, on users' opinion polarity and engagement behaviours to investigate whether AI can trigger these cognitive biases.

2.2 Applications of AI-curated Information in Social Media

Prior applications of AI-curated information in social media are largely motivated by the need to address mis/disinformation through fact-checker AI summaries [9, 37, 39, 50], or to pop 'filter bubbles' [14, 66, 72, 120, 120]. Filter bubbles refer to the personalised curation of online content, such as recommended Reddit posts or YouTube feeds, tailored to the users' perceived interests to foster engagement [72]. However, these filter bubbles can result in online radicalisation rabbit holes [25]—including moderate to extreme fringe subculture radicalisation pipelines on platforms such as YouTube [27, 52]. Moreover, platforms target engagement and retention, thus platforms may cater by sorting comments by controversial or heated posts, or alternatively by the majority 'most-liked' consensus as seen on Reddit (via upvotes) and YouTube (by likes). Sorting by controversial posts for engagement risks enhancing polarising in and out-group dynamics, where users assume a group identity framed around antagonising an opposing 'out-group' [105]. Conversely, sorting comments by most liked reasserts the majority consensus, potentially reinforcing the spiral of silence online.

Thus, AI summaries provide an opportunity to highlight narratives that may be buried due to majority opinion, brigading (where outside users try to overtake a forum/thread to manipulate the perception of the majority consensus [71]), or shadowbanned/delisted due to opaque social media algorithms [40]. As such, we examine whether AI summaries of users' perspectives in a Reddit thread could help reduce opinion polarisation by providing objective (rather than emotive) analyses of users' arguments, alongside presenting the proportion of commenter agreement in a thread rather than leaving it up to a subjective/biased user's perception.

Current AI-based tools demonstrate various strategies to help users navigate their exposure to different viewpoints in a filtered and algorithm-curated social media landscape. For example, the browser extension, Pop, augments 𝕏 (formerly Twitter) feeds with news tweets from agencies of differing ideological stances, though its effectiveness in helping users learn about different perspectives compared to manual searching remains empirically unevaluated [66]. Meanwhile, the web application Social Mirror encourages users to follow accounts with contrasting political ideologies to reduce participants' perceptions of a politically homogeneous network and thus prevent their feed from becoming an algorithm-driven political echo chamber [74]. However, these approaches encourage users to seek *subjective* sources, rather than relying on an external AI summary of opinions/thoughts—which could provide a more objective detoxified analysis of a heated topic.

Solutions to help reading audiences depolarise their position online currently rely on AI agents *actively partaking* in the online discussion—such as via AI-driven mediator chatbots [28], or multi-agent personas providing different perspectives [120]. Thus, our study investigates whether a *passive* approach of providing detoxified AI summaries alone would depolarise users without having to alter the actual online discussions through interactive chatbots as seen in related work [28, 120].

3 Method

The core objective of this study is to identify how AI summaries of *narratives* and the *percentages* of agreement/disagreement influence a social media user's opinion-making processes and their experience regarding their willingness to engage in a social media thread. Here, we break down each of the design components of the study starting with the thread topics and discussion design. Thereafter, we outline the design considerations of different summarisation tools—which mimic existing implementations by Meta (for Facebook) [60], 𝕏 [68], and Ground News [14, 31]. Finally, we present the quantitative and qualitative measures used in our study and the procedure followed in our experiment. We note that our simulated topics and data collection process occurred in April 2025, reflecting current events.

3.1 Topic Selection and Polarity

We replicated a Reddit-style social media platform, centred on a parent post on a given topic where users can comment and (sub)reply. This threaded format reflects real-world platforms such as Reddit, Facebook, and 𝕏, all of which are currently exploring AI tools for summarising conversations and visualising the proportion of differing opinions [66]. For the overall topic, we simulated polarised comments across six threads using Large Language Models (LLMs) on a custom site stylised to look like Reddit and accessible to the consenting participants only.

We divided these threads into two categories: *civil* and *toxic*. We constructed the civil threads to have conversations that rely on deliberative *constructive* discourse—where comments must be

evidence-driven or rely on logical argumentation without relying on fallacies, personal attacks, or offensive language [25]. Conversely, toxic threads were constructed to rely on extreme polarity whereby users may also rely on evidence and logic, but contain *destructive* discourse strategies such as personal ad hominem attacks, logical fallacies, and extreme disregard for opposing views.

We select our topics to offer a salient spectrum of examples of current events topics with high-probability of toxic discourse (i.e., high affective polarisation, which correlates to toxic discourse online [25]), to low-stakes topics, which are more likely to contain civil discourse due to their low affective polarisation (i.e., disdain for the other 'out-group' [105]). We select our topics with toxic discourse inspired by real topics/discourse on Reddit in April 2025 using the controversy/polarisation-driven top 'hot' threads filter on r/politics (archive of the subreddit topics [78, 79]). For our civil threads, we target the Fukushima wastewater treatment debate for its international focus [80, 113], while remaining less familiar to a US audience (reducing likelihood for toxic affective polarised discourse) [27], and public transport [81, 102, 103] and social media under-16 ban topics (also in current events discourse [77, 101]).

The civil threads include the following topics:

(1) Fukushima (FUKU)—with a post regarding the proposed discharge of treated radioactive water from the Fukushima Daiichi Nuclear Power Plant into the Pacific Ocean.
   (a) The parent post on the simulated subreddit r/science makes the claim "The Japanese government should discharge the treated Fukushima nuclear wastewater into the Pacific Ocean rather than to store it onsite."
(2) Free Public Transport (PT):
   (a) The parent post on the simulated subreddit r/transport makes the claim that "Public transport should be free across US cities."
(3) Social Media Under-16 Ban (SOC):
   (a) The parent post on the simulated subreddit r/social makes the claim that "Social media should be banned for those under the age of 16."

The three toxic threads include the topics:

(1) Ukraine-War Peace Deal (UKR)—focusing on the February 2025 news on an immediate ceasefire in Ukraine proposed by US President Donald Trump.
   (a) The parent post on the simulated subreddit r/politics makes the claim "Trump is doing the right thing by pushing for an immediate ceasefire between Ukraine and Russia."
(2) Elon Musk's role in the Department of Government Efficiency
   (a) The parent post on the simulated subreddit r/politics makes the claim that "It is a good thing that Elon Musk is in-charge of the new Department of Government Efficiency (DOGE)."
(3) Gun Control (GNC)—focusing on a proposed ban on Military Style Semi-Automatics
   (a) The parent post on the simulated subreddit r/politics makes the claim that "Military-Style Semi-Automatic Rifles Should Be Banned in the U.S."

All topics follow the format shown in Figure 1. Each post has ten three-post comments for consistency, where every 1st level parent comment has two visible replies (as Reddit by default shows first-order comments before hiding nested comments under a 'Show More' button). We utilised this total of 30 comments to reflect small threaded conversations so that the participant can observe civil and toxic *argumentation* between commentators/repliers, and to approximate the median number of comments seen across all Reddit threads [109]. Likewise, we prioritised the first and second level posts (comments and replies) over long nested comment chains, as the majority of Reddit posts involve the first or second level comments, with the frequency of comment changes of n-length decreasing logarithmically [109].

Participants read all six topics, where we manipulated whether they see the textual *AI narratives* summarisation tool, the *AI percentages* tool, or neither (*control*). We display an exact excerpt from the Ukraine topic thread with both the *AI narratives* and *AI percentages* tools in Figure 1. Across the civil and toxic thread topics, we manipulated the visualised percentages of agreement in the comments themselves to mimic degrees of polarisation on threads, consisting of a 'balanced' thread where ~50% of comments agree with the premise and ~50% disagree ('balanced' condition), a strong pro-topic echo chamber thread where ~90% agree with the topic premise and ~10% disagree ('polarised-agree' condition); as well as an anti-topic echo chamber thread where ~10% agree with the topic premise, and ~90% disagree ('polarised-disagree' condition).

These three proportions (balanced, polarised-agree, polarised-disagree) were applied to all topics, generating 18 thread transcript variants: 6 topics * 3 agreement proportions (balanced, polarised-agree, polarised-disagree). These transcripts were counter-balanced and each participant saw one transcript from each topic. For example, one participant saw one of the civil topics with ideologically balanced comments, another with strong disagreement, and one with strong agreement; but did not see two civil topics with the same proportion (same for the toxic threads). This ensured that we could observe the effects of AI tools in the cases of echo chambers and cases where debate is balanced to determine how visualised percentages of commenter agreement/disagreement could influence social conformity, with users conforming or moderating their stance towards the majority group of commentators.

3.1.1 *Comment Design.*

To ensure consistency in generating clear toxic or civil comments that agree or disagree with a thread's premise, we utilised GPT-4.5 to generate our comments [69]. Utilising AI to simulate online discourse is common in HCI research [26, 54, 73], and it enabled us to control the polarity (disagreement/agreement) as well as the toxicity of the comments. Likewise, simulating posts for AI-driven polarisation studies resolves the ethical issues of running manipulation experiments on live unknowing users who have not consented to a study and are not debriefed after the experiment [98].

We specified an average post length of 50–100 words to mirror the posts on related subreddits [7, 109]. GPT-4.5 was specifically chosen for its state-of-the-art argumentation quality and realistic human-like speech patterns as of April 2025 [69]. Using a single, high-performing AI model ensures consistency in argumentative

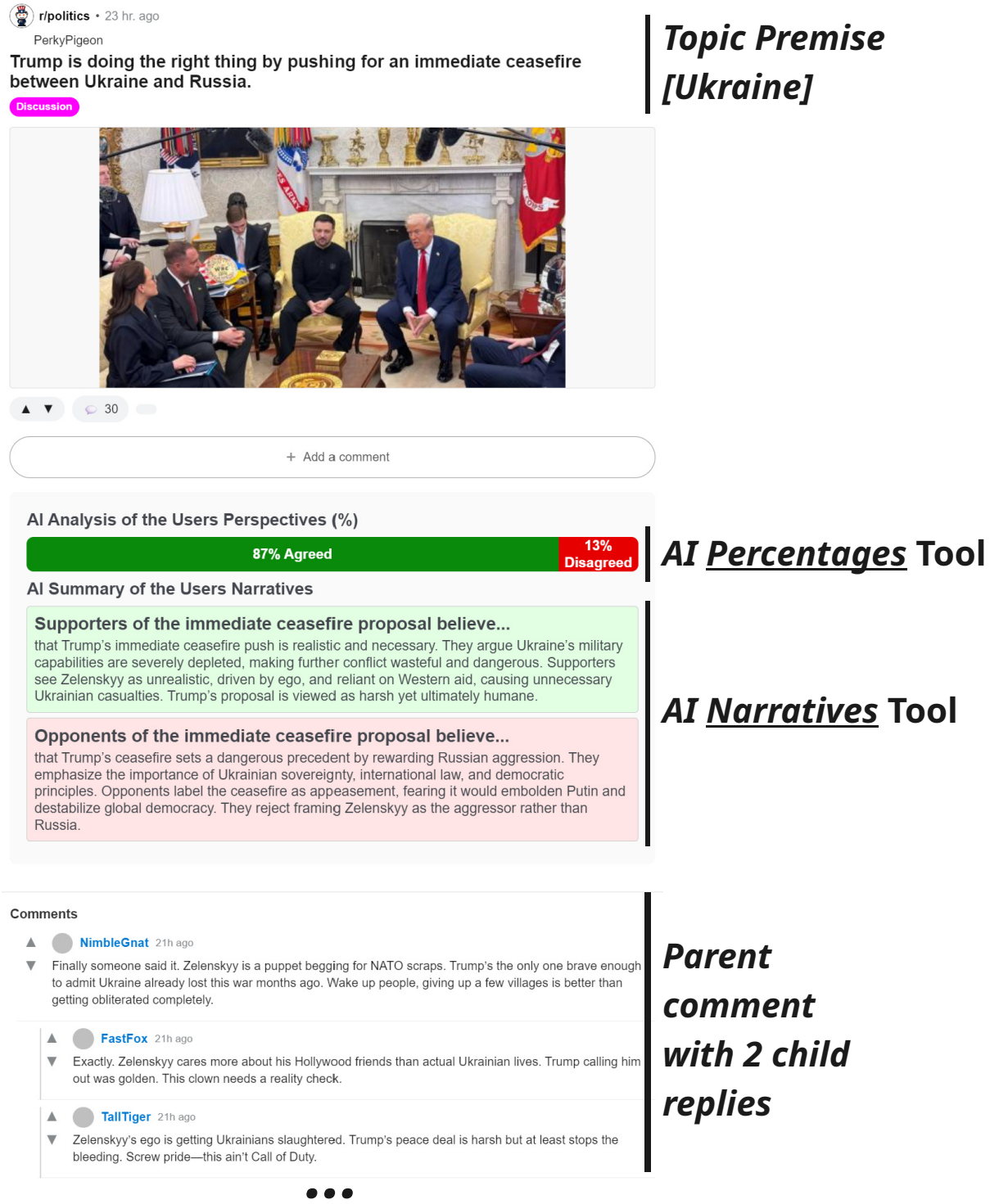


Figure 1: Example of the simulated Reddit thread transcript with both the *AI narratives* summary and the *AI percentages* tools, and one of the ten comment chains (with two replies each). The remaining 27 comments (9 parent comments with 2 replies each) are hidden for clarity.

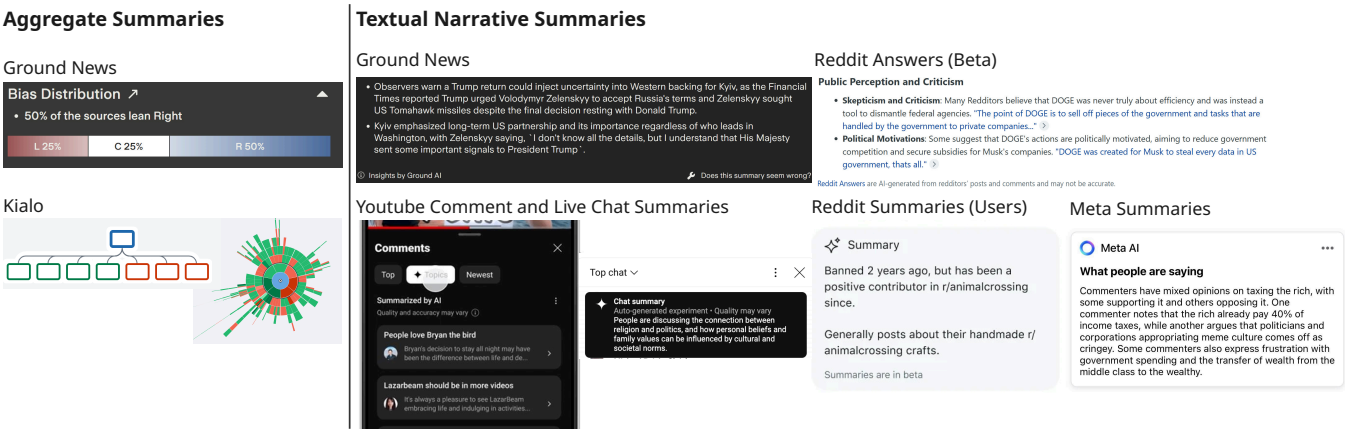


Figure 2: Examples of aggregate 'agreement' summaries—such as political alignment (Ground News [31]), and Kialo's visualisations [45]. Summaries can also summarise text to provide a narrative to the conversation/comments, such as Ground News [31], Reddit Answers [82], and the comment-section summaries on Meta's social media platforms [60].

quality and sentiment across different viewpoints, avoiding potential biases where one side of a real Reddit thread might inadvertently have stronger or weaker arguments. Beyond neutral summaries, generative AI can also mimic emotive toxic political discourse [26], as well as informal Reddit-like speech and humour [24, 100]. Our prompts instructed GPT-4.5 to generate 10 comment responses with 2 replies, where we simulated *civil* (evidence-driven, logical chains-of-thought, relevant) or *toxic* (ad hominem, logical fallacies, destructive) comments based on the topic. We provide all prompts to generate our comments and AI narrative summary text across all six topics in Appendix A.1–A.2 for reproducibility.

3.1.2 *Summaries Design.*

We selected our AI summary modalities based on established examples of visualisation and textual summarisation designs used in existing platforms, with examples from the news-aggregator Ground News [31], Meta ecosystem (Facebook, Instagram, WhatsApp, Messenger, and Threads) [60, 95], 𝕏 [68], and Google (YouTube [117], and their 'AI Overviews' for web searches [83]). We observed two main types of AI modalities: visual/numeric *aggregators*—which provide breakdowns of consensus and agreement, and *narrative summaries*—which summarise chats, or webpages to give breakdowns of perspectives while remaining agonistic to the level of agreement. Platforms like Kialo visualise agreement in the threads (Figure 2), while platforms like YouTube and Reddit show rely on non-AI likes/karma from users themselves to visualise agreement directly from the users. Nonetheless, AI are replacing these human-driven approaches, with research in aggregate numeric summaries including bias/agreement indicator labels and misinformation visualisations [9, 37, 68] (including for source bias (Figure 2)), with optional descriptions to provide context [46, 95].

We used the real number of comments which agree and those that disagree to generate our 'AI-generated' agreement/disagreement percentages—which allows us to identify if participants perceive the 'AI' percentages to be biased (when in reality, they are the ground truth). We used the selected batches of agreeing/disagreeing comments to generate the AI narrative summaries—which target ~50 words in length to mimic Facebook comment section narrative summaries [60]. We also targeted an emotionless neutral tone for our summary prompts as the tone of speech online can influence a user's engagement and their opinion on a discussion topic [2, 118].

In our simulated subreddits, we disabled Karma (i.e., visible upvotes/downvotes, as common in political subreddits), gave anonymous Reddit-like usernames which are topic-agonistic (e.g., QuantumQuokka, YummyYak), and disabled profile pictures for consistency and to reduce potential biases [35, 94, 106], with comment examples visualised in Figure 1. We set all comments to the same timeframe ("21 hours ago") to reduce temporal recency biases.

3.2 Measures

To address RQ1's focus on opinion *change* and polarity perceptions on the social media thread, we measured participants' initial opinion and final opinion through a 7-pt Likert scale on agreeing/disagreeing to the topic's premise (e.g. "I support a ban on Military-Style Semi-Automatics"), an approach utilised in related work on online polarisation [5, 6, 27, 28].

We also measured participants' *perceived* topic polarity, defined as their estimate of the proportion of comments that agree or disagree with the topic's premise. While we expect that providing AI-generated percentage estimates will improve their accuracy, we are equally interested in how an AI-generated summary alone might affect their perception of the thread's overall polarity. For instance, if users are shown two equally sized summaries—one for, one against the premise—they might infer an even split of opinions, even if the actual distribution of comments is strongly one-sided.

This measure of *perceived* polarity is also valuable in the control (no-AI) condition, as it allows us to see how accurately people gauge the distribution of opinions when relying solely on their own judgement, especially if the participants hold strong views on the topic, as this would align with the theory of *Pluralistic Ignorance*.

For RQ2, we investigate participants' desire and willingness to engage with a thread after reading it. We define engagement

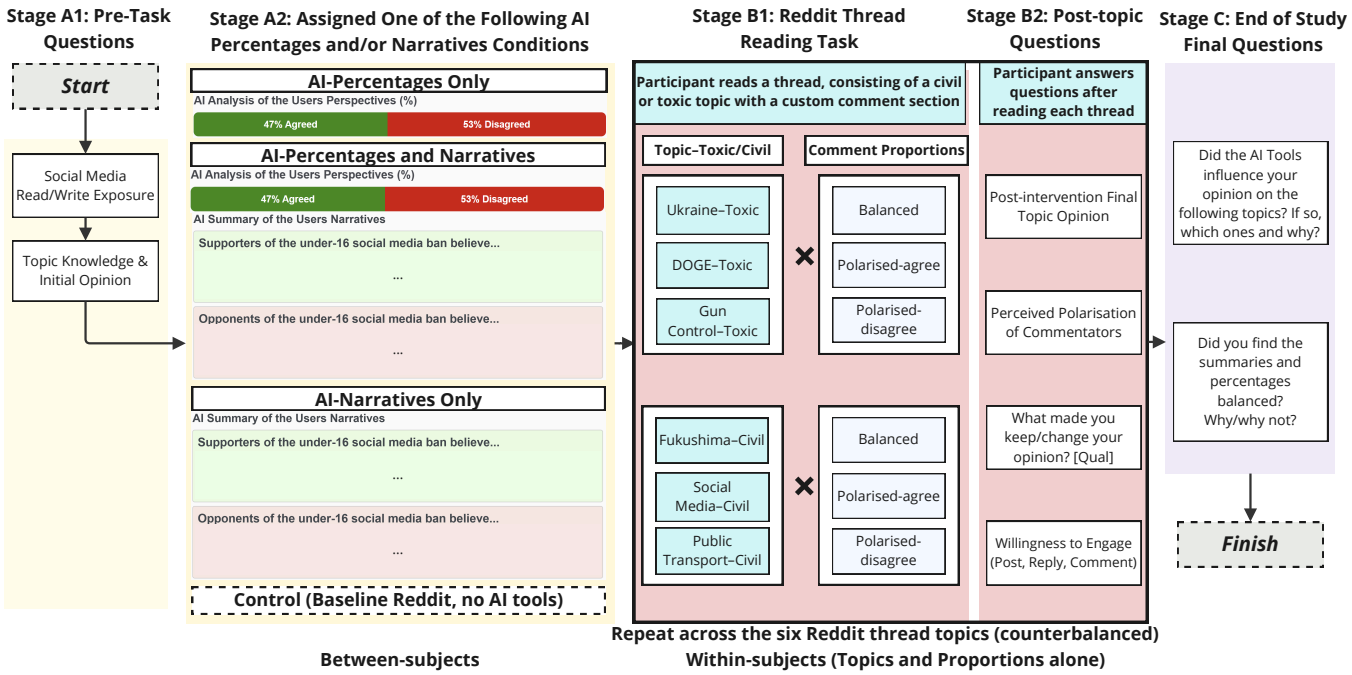


Figure 3: The experimental flow of our study—covering demographic, predictor and pre-test questions in Stage A1, their assignment of their AI summary tool condition for the entire experiment in Stage A2, and the process loop of reading a Reddit thread in Stage B1 and answering questions in Stage B2. This latter process repeats for each of the six counterbalanced Reddit threads before concluding with two qualitative questions at Stage C.

as whether they would want to write a comment, cast an upvote/downvote, or reply to someone else's comment. This approach tests the *Spiral of Silence* theory, which predicts that individuals are less inclined to respond when they are in the (overwhelming) minority [34, 64]—most clearly demonstrated in conditions where only a small minority of comments support their stance. We measured this *'willingness to engage'* tendency using a 7-point Likert scale for both internal consistency and for mixed-model analysis. This allowed participants to indicate whether they *would* respond, without necessarily disclosing *how*. Although this method has its limitations, it avoids the risk that participants self-censor or rush their written comments, improving the validity of the engagement measure.

Thus, our measures for participant opinion change (RQ1) and their willingness to engage in the thread (RQ2) include:

- RQ1 Opinion Measures:
  - Initial Opinion Agreement Likert—a 7-point Likert scale on how much the user agrees/disagrees with the topic premise.
    - Example: "Military-Style Semi-Automatic rifles should be banned in the U.S."
  - Final Opinion Likert—same Likert as above but asked after reading the Reddit thread.
    - "After reading the Reddit thread, how much do you agree or disagree with the statement that: <Premise>?"
  - Final Opinion Qualitative Question:
    - "Please explain your decision above. What made you keep or change your opinion on the topic?"
  - Perceived Polarity Slider:
    - "On a scale from 0-to-100, where 0 means 'no comments supported' and 100 means 'all comments supported', how much support did the <Topic Premise> receive in the Reddit thread?"
- RQ2 Engagement:
  - Willingness to Engage Likert (7-point scale):
    - "After reading the Reddit thread, how likely are you to engage/respond to this Reddit thread (including replying, upvoting/downvoting comments, etc.)?"
- Open-ended Questions Asked at the Conclusion of the Study:
  - "Overall, what are your thoughts on the AI summarisation tool?"
  - "Did you find the AI summarisation tool fair and balanced across the topics?"

3.3 Procedure

We utilised simulation-based power analysis to calculate our minimum sample size. Our simulation is based on effect size findings from previous research on the effect of AI-driven mediators on user opinion, polarisation, and behaviour [28], the role of AI contextual bots like nudging tools [9] and the impact of social conformity on perceptions of comment sections [110]. We calculated a minimum sample size of 140 to achieve a power of 0.8 following established

methodology recommended for psychology studies by Cohen [16], as standard for HCI experiments [27, 53, 70, 85, 116].

We used the Prolific crowdsourcing platform to recruit 144 participants (72M, 72F) to ensure an equal distribution of participants across all conditions. All participants are from the United States with a minimum approval rate of 98%, and with an equal split of self-reported Republicans, Democrats, and Independents for political balance. We sample only first-language English speakers due to the nature of our experiment.

To minimise the risk of participants' being aware of the experiment's aim, we masked the study's true purpose in the following ways. Firstly, we framed our study as an opinion polling exercise, highlighting our aim to "understand perspectives on Reddit threads" and instructed participants to browse as they normally would. Furthermore, each participant only saw one AI condition throughout the experiment (AI percentages only, AI narrative summaries only, both, or neither) to prevent any changing layouts from priming them into focusing on the AI tools. Finally, we included a buffer task between each thread-reading task where participants wrote their thoughts on the political topic to contribute to the public polling guise (akin to related opinion polling and social media conformity studies [43, 110]). We exclude any mention of the AI tools in our questioning until the conclusion of the experiment (i.e., after collecting all RQ1/RQ2 quantitative measures).

We visualise our full 30-minute survey in Figure 3, which was approved by our university's Human Ethics Committee.

Stage A consisted of the initial Plain Language Statement, the consent form and the demographic questionnaire, where we collected how often each participant read social media (Never, Sometimes over: the past six months, ...past month, ...weekly, ...daily), how often they post and actively partake in social media discussions (same scale as the prior question), their initial opinion on each of the six thread topics (1-to-7 Likert agreement scale), and their self-reported familiarity/knowledge (1-to-7 Likert scale). We then assigned them to one of four AI conditions across all six Reddit threads: AI percentages only (where participants see a percentage bar of how many commentators *overall* agreed or disagreed with the topic's premise), AI narratives only, both, or a control (i.e., standard Reddit post with comments, no AI).

Stage B reflects the core experiment where participants have up to three minutes to browse our interface to read the post, (if applicable) AI tools, and the comment section. We opted for three minutes per thread before prompting the participant to wrap up to maintain consistent payment/timing as well as due to our internal pilot testing identifying that participants took approximately two minutes to read all comments, thus allowing a comfortable maximum time for those that tend to engage in 'deep reading' while allowing participants to proceed early if they are skim readers. As aforementioned, participants read one thread at a time, consisting of a topic (toxic or civil) where the comments either reflect balanced argumentation or one-sided echo chamber polarity (polarised-agree or polarise-disagreement condition). After the participant completed their reading task, they completed the quantitative and qualitative measures from Subsection 3.2 (no time limit). Stages B1 and B2 then repeated across the six thread topics, where the topics and proportion of comment polarity differed within-subjects, while the presence of the AI percentages and/or AI narratives tool differed between-subjects.

After all six topics, participants completed the exit questions in Stage C, where we probed their experience and perceptions of the AI tools. Finally, we gave participants a debrief document explaining the real purpose of the experiment (understanding conformity from comments and AI summariser tools), and references to reputable articles on the topics (e.g., International Atomic Energy Agency for Fukushima [36] and the Associated Press for the Ukraine-war [48]) as recommended for responsible polarisation studies and to reduce inadvertent polarisation of the participants [27].

4 Results

We analyse the results from our 144-participant sample using a mixed-methods approach consisting of Generalised Linear Mixed Models (GLMMs) and Cumulative Link Mixed Models (CLMMs) for quantitative analysis, alongside inductive thematic analysis for our open-ended qualitative questions. We calculated the Variance Inflation Factors (VIF) for all RQ1/RQ2 mixed models to assess for multicollinearity among the independent variables. All VIF values were below five, indicating no significant linear dependence among the predictors [76].

We analyse RQ1's focus on how these AI tools could shape user opinions on the discussion topics through two lenses: the *direction* of opinion change (towards extremes or towards centre/neutral view), and its *magnitude*; as visualised in Figure 4 with significant effects shown in Table 1.

We measure opinion change (RQ1) as the shift in opinion from the pre-experiment 7-point Likert scale on agreement for each of the topics, compared to their scores after reading each topic with or without the AI narratives and/or AI percentages summarisation tools. We capture the magnitude of opinion change through the absolute difference between pre- and post-test opinion Likert scales as used in related work [6, 27].

We measure the magnitude of opinion change as both the relative Cohen's d effect size from the AI tools and the absolute opinion swing. For relative effect size changes, a negative value indicates that the intervention had a reduced opinion change magnitude compared to the overall marginal mean opinion change, while positive effect size values indicate a higher than marginal mean effect size change. Where effect size error bars intersect 0 (e.g., as seen in Figure 5), this indicates that there is no significant effect. We also discuss the estimated marginal means (emmeans) values in Likert values themselves to give the expected mean opinion change for each intervention group.

We also measure the *direction* of opinion change to observe how comments that are polarised in either political direction with/without the AI tools could influence whether they become more or less polarised. For instance, a user may have a significant opinion swing from 'Agree' to 'Somewhat disagree' (a 3-point Likert shift), which indicates a strong opinion change effect but not necessarily a risk of extreme polarisation or radicalisation. Conversely, a user may shift from 'neither agree nor disagree' to 'strongly disagree', which is also a 3-point *magnitude* swing, but this indicates that the comments and/or tools had a *directional* and *radicalising*

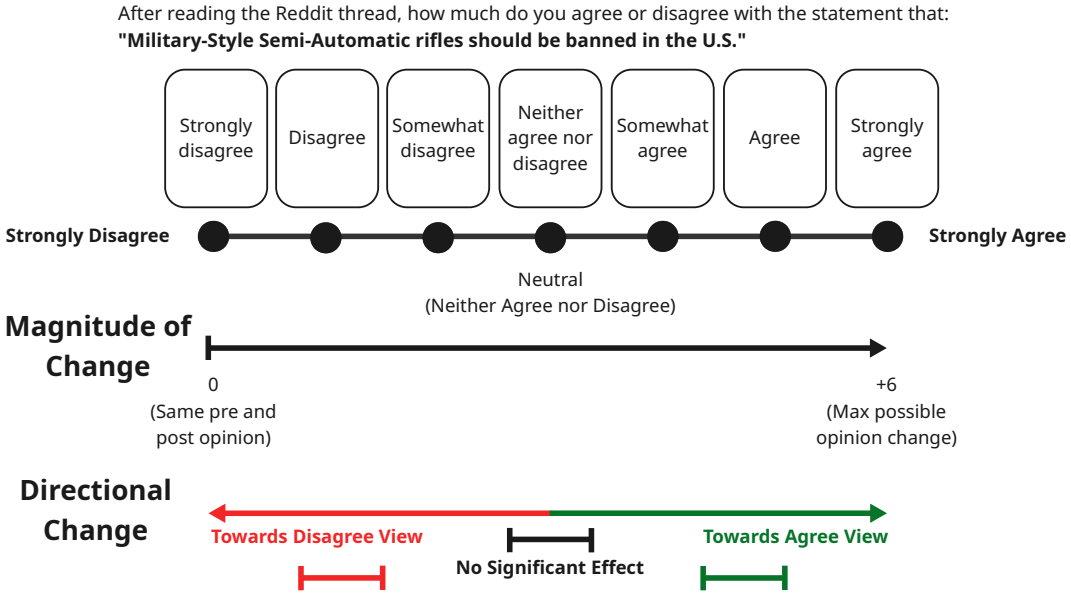


Figure 4: The two approaches for measuring opinion change (RQ1), consisting of opinion swings towards making their positions more extreme (*directional*), or exacerbating their overall opinion change (*magnitude*).

Table 1: RQ1-M = Magnitude of Change Mixed-Model, indicating the significant predictors and the AI tools for enhancing the magnitude of opinion change towards the majority (or balanced centre) view; RQ1-D = Directional Mixed-Model, indicating predictors/tools relevant for which direction (towards pro, or anti, or centre) view; RQ2 = Willingness to Engage after reading the thread mixed-model, indicating whether the predictors made participants more or less willing to engage in the thread.

	Significant Predictors							Significant AI Interventions		
	Initial Opinion	Topic Knowledge	Social Read	Social Write	Opinion Congruence	Comment Proportion (Directed)	Thread Toxicity	Percentages Only	Narratives Only	Both
RQ1-M	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓
RQ1-D	✓	✗	✗	✗	✓	✓	✗	✓	✗	✓
RQ2	✓	✓	✗	✓	✓	✓	✗	✗	✗	✗

effect. We also investigate whether participants' opinions moderated (RQ1) towards the centre 'neither agree nor disagree' stance, as users' opinions may not necessarily polarise to the anti or pro stance views, but instead *depolarise* towards the neutral stance. For RQ2, we analyse participants' willingness to engage (respond, comment, upvote/downvote) in each thread through a single CLMM with the findings also summarised in Table 1.

Finally, we conclude with our three qualitative questions (Figure 3) as analysed via the general inductive approach [104]. The main author began by reviewing the qualitative data to deepen their understanding of its content. This was followed by the identification of a preliminary set of categories, which were then collaboratively refined with another member of the research team through an iterative process. Once the categories were finalised, both researchers independently used these codes to systematically analyse the participants' responses in a deductive manner.

4.1 Quantitative Findings

Overall, we found that the AI percentages, especially when combined with the AI narratives summaries, significantly increased both the direction and magnitude of participants' opinion shifts toward the majority view. Likewise, we found no evidence of widespread belief reinforcement (i.e., 'doubling down') when participants encountered opposing comments. The inclusion of AI percentages amplified conformity (Cohen's d = 0.81) and led to greater opinion change (up to a 1.5 Likert swing) compared to the summaries alone (0.9 swing). Participants' *perceived* agreement in the thread also played a key role: when participants rated the discussion as balanced, they moderated their views, whereas perceived majorities intensified polarisation. Notably, the AI narratives tool on its own often led participants to wrongly perceive debates as balanced even when they had the strongly polarised comment threads, which reduced their opinion swings.

For RQ2, the AI tools had no significant effect on their willingness to engage. Instead, engagement was more strongly influenced by

the thread's tone: participants were more willing to engage in threads with civil discourse, even when in the minority, while threads with toxic discourse discouraged participation, especially for those whose views were in the minority.

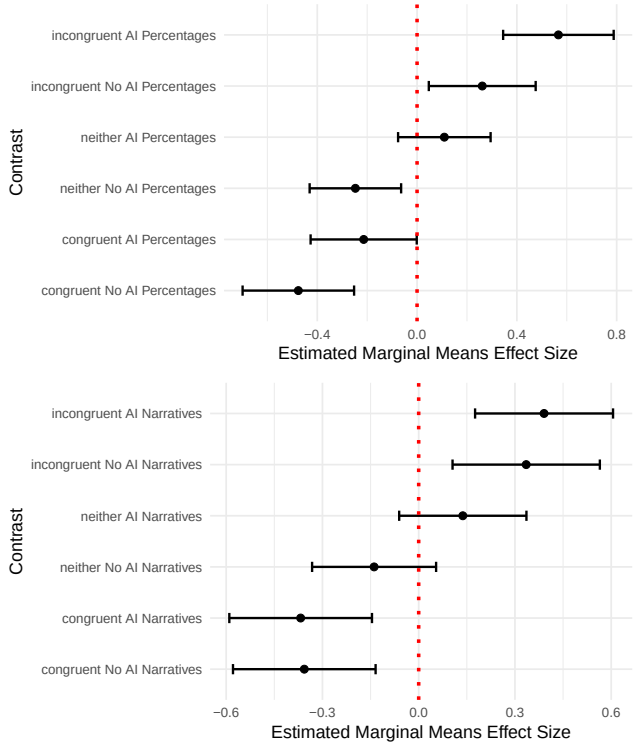


Figure 5: Emmeans Cohen's d effect size for the magnitude of opinion change based on whether the majority of comments agreed with the participant's stance or not (congruence), given the presence of the AI *percentages* (top plot) and/or AI *narratives* (bottom plot) tools. Incongruent refers to a participant having an opposing view to the majority of commenters, congruent refers to them sharing the majority opinion, and neither is where opinions are balanced. Negative values represent that the magnitude of change is below the marginal mean of the overall mixed-model, while positive represents that the magnitude of change is above it.

4.1.1 *RQ1—Magnitude of Opinion Change from the AI Summaries.*

Our analysis reveals that the different AI tool conditions produced distinct effects on the magnitude of participants' opinion change when compared to the control without-AI group (Figure 5). The most substantial impact came from the combined use of the tools (Figure 6). The condition with both *AI percentages* and *AI narratives* produced the greatest opinion swing, with an estimated marginal mean (emmeans) shift of 1.5 Likert points (SE = 0.13, $p < 0.01$).

In contrast, the *AI narratives* summaries alone had a moderating effect. This condition resulted in the smallest opinion swing of 0.9 Likert points (SE = 0.13, $p < 0.01$), thereby reducing the amount

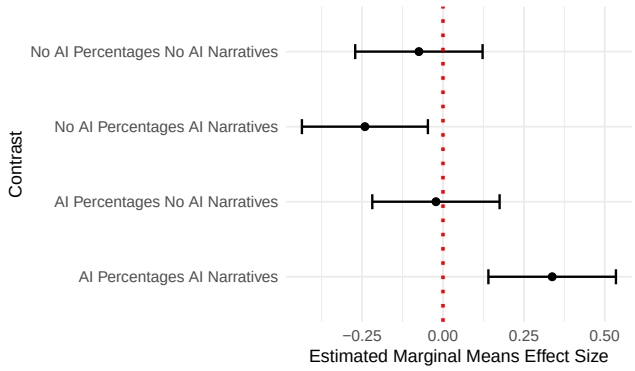


Figure 6: Emmeans Cohen's d effect size for the magnitude of opinion change overall based on the presence (or lack) of the *AI narratives* and/or *AI percentages* tools. Values represent magnitude of change above/below the marginal mean.

of opinion change to a level even below the control condition's comment-section only 1.089 marginal mean opinion swing (SE = 0.13, $p < 0.01$), reinforcing that comment sections alone can have a conformity effect.

The *AI percentages* tool was the primary driver for increasing the magnitude of opinion change. A direct pairwise comparison shows that adding the *AI percentages* tool to the *AI narratives* tool (i.e., comparing the Both condition to the *AI narratives only* condition) also significantly increased the magnitude of the opinion swing (Cohen's d = 0.47, SE = 0.13, $p < 0.01$).

Significant predictors for the magnitude of opinion change included prior topic knowledge and social media posting habits, as well as the thread's toxicity. Participants with higher self-reported knowledge were less likely to change their opinion (Cohen's d = 0.53, SE = 0.22, $p < 0.05$). Participants who rarely post on social media were more likely to change their opinion compared to daily posters (Cohen's d = 0.51, SE = 0.24, $p < 0.05$).

The effect of the overall thread's toxicity also affected opinion change, whereby participants who agreed with the majority consensus of a civil topic were far more likely to change their opinion compared to a toxic thread where the majority of participants disagreed (Cohen's d = 0.66, SE = 0.15, $p < 0.01$). A participant's *perceived* agreement of the thread (i.e., the percentage of agreement and disagreement they believe is in the thread, as measured after reading the thread) also had a significant effect, where perceptions of one-sidedness increased the magnitude of opinion swing (Cohen's d = 0.280, SE = 0.083, $p < 0.01$). While this effect considers controversial topics with toxic discourse (i.e., toxic threads), we also investigated how the topics themselves differed within their respective civil (Fukushima vs. Public Transport vs. Social Media U16 Ban) and toxic discourse (DOGE vs. Ukraine vs. Gun Control) threads. We observed no significant effects between these topics within their respective civil or toxic category.

4.1.2 *RQ1—Direction of Opinion Change from the AI Summaries.*

Next, we analysed the direction of opinion change to determine if the AI tools pushed participants towards the pro- or anti-topic stance (Figure 7). We found a strong conformity effect towards the

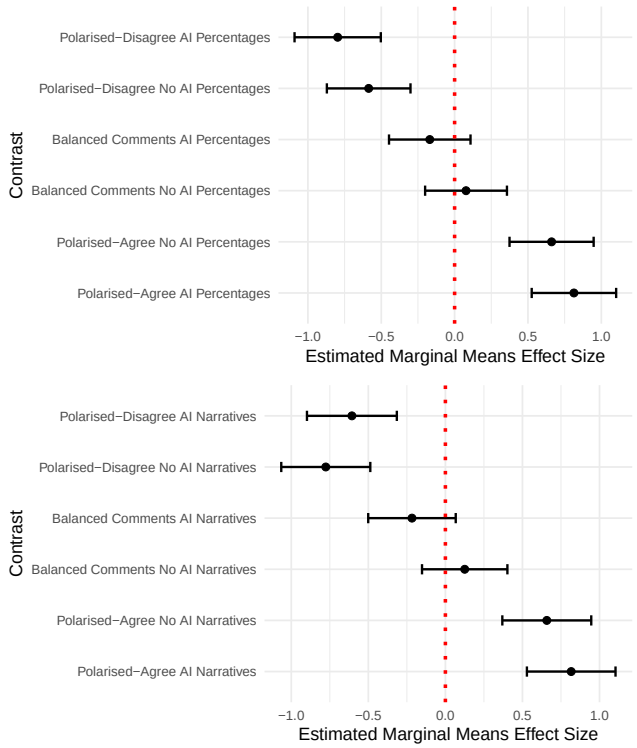


Figure 7: Emmeans Cohen's d effect size for the direction of opinion change based on the comments stance and the presence of the *AI percentages* tool overall (top plot), as well as the *AI narratives* tool overall (bottom plot). In these *directional* plots, a negative effect size indicates an opinion change towards the anti-topic stance, while a positive effect size indicates opinion change towards the pro-topic stance.

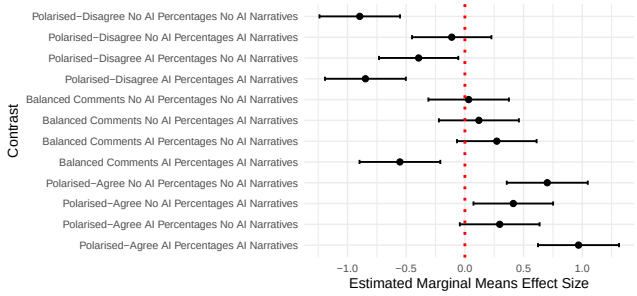


Figure 8: Emmeans Cohen's d effect size for the direction of opinion change based on the comments stance and the the presence (or lack) of the *AI narratives* and/or *AI percentages* tools.

majority polarised threads which was significantly amplified by the *AI percentages* tool, and further amplified when paired with the *AI narratives* tool. For instance, in threads where the majority of comments disagreed with the topic premise ('polarised-disagree'), the *AI percentages* tool induced a strong opinion change in that direction (Cohen's d = 0.62, SE = 0.12, p < 0.01), resulting in a marginal mean final opinion of 3.60/7 (SE = 0.20, p < 0.01). In threads where the majority agreed ('polarised-agree'), it pushed participants toward agreement (Cohen's d = 0.67, SE = 0.15, p < 0.01), with an emmeans final opinion of 4.85/7 (SE = 0.20, p < 0.01).

The combination of both AI tools further amplified this directional shift. When both tools were present, the opinion swing towards the majority view increased by 0.67 Likert values compared to the control (SE = 0.24, p < 0.05). The *AI narratives* tool alone did not produce a significant directional effect. The most significant predictor of the directional change in opinion was the initial opinion of the participant on the topic (Cohen's d = 0.54, SE = 0.13, p < 0.01).

Notably, while the *AI narratives* tool reduced the overall *magnitude* of opinion change, we did not observe any directional swing either towards pro/anti-topic stances nor towards the central 'neither agree nor disagree' stance. This indicates that the tool had a moderating effect on just the *magnitude* of opinion change, rather than a depolarising effect towards a neutral viewpoint. Moreover, no condition resulted in a significant depolarisation effect (i.e., an opinion change *towards* the centre neutral view).

Overall, we found that opinion change was primarily driven by social conformity towards the majority opinion, supporting previous research showing that commenter consensus can influence readers' opinions [110]. Notably, social conformity towards the majority opinion was significant for the control No AI condition, confirming that participants read the comments (Figure 8, where this effect was enhanced (AI percentages) or weakened (AI narratives alone reducing the magnitude of change) by the AI tools, highlighting that participants read the AI tools as well.

4.1.3 *RQ1—Perceived Polarisation of the Commenters.*

After reading each thread, we asked the participants to rate their percentage of comments that they believed agreed or disagreed with the topic premise on a scale of 0-to-100.

We model the participants' perceived polarisation measure through an ordinal beta regression model, which allows for bounded numeric 0-to-100 scales [47]. We output the density plots and distribution of the participants' perceived polarisation of the threads in Figure 9, noting that a perfect recall/estimation of the thread polarisation should be at the ~90% (+/-3%) agree, 13% (+/3%) agree and 50% agree (+/-3%) to reflect our three proportions (polarised-agree, polarised-disagree, and balanced threads).

We found that the *AI narratives* tool alone led to a notable misperception. Because it presented both arguments, participants often perceived the discussion as balanced (~50% agreement), even when the actual comments were predominantly polarised (Figure 9). This misperception helps explain the tool's moderating effect: when participants *perceived* a thread as balanced, their opinions shifted by a 0.57 marginal mean opinion swing (SE = 0.24, $p < 0.05$)—a significant change compared to the overall marginal mean of opinion 0.74 (SE = 0.47, $p = 0.12$).

By contrast, the *AI percentages* tool appeared to amplify conformity. Unlike *AI narratives*, it provided an accurate depiction of the real majority consensus (Figure 9), which in turn led to greater conformity and larger shifts in opinion.

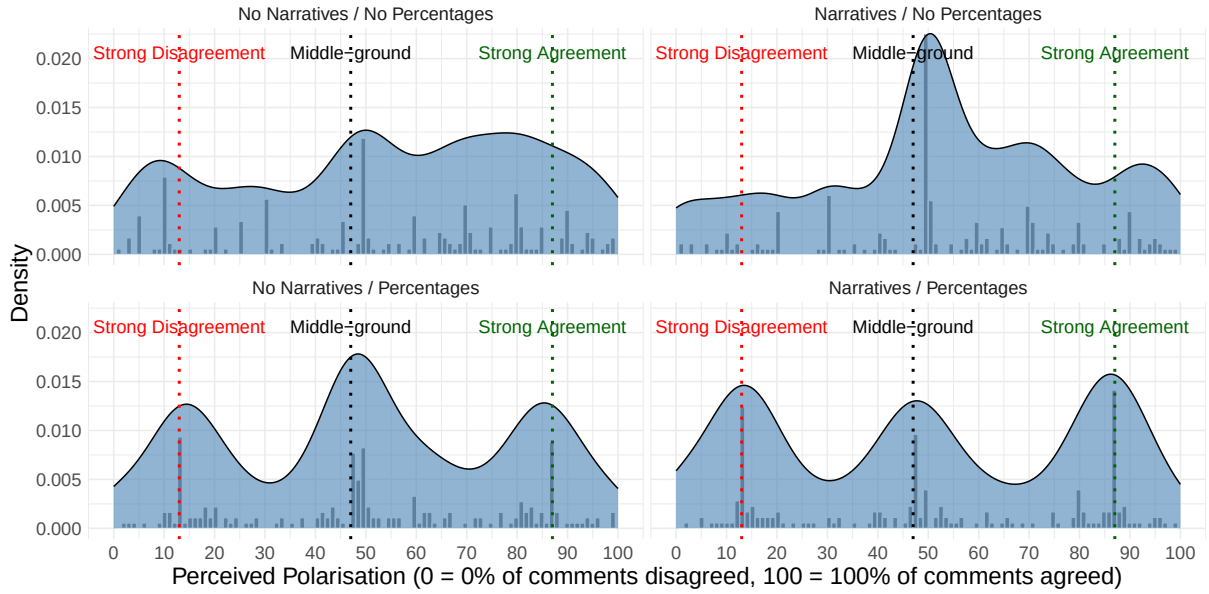


Figure 9: Perceived Polarisation by *AI narratives* and *AI percentages* Conditions.

4.1.4 *RQ2—Willingness to Engage.*

For our second research question, we found that the AI tools had no significant effect on participants' willingness to engage with the discussion threads (i.e., comment, reply, or vote). Instead, the primary driver of engagement was the civility of the discussion.

Participants were significantly more willing to engage in civil threads (e.g., Public Transport, Social Media Ban) than in toxic ones (e.g., DOGE, Gun Control, Ukraine War), with a moderate effect size (Cohen's d = 0.54, SE = 0.10, $p < 0.01$). This finding suggests that a constructive environment is a more critical factor for participation than the perception of consensus.

Furthermore, we observed an interaction between toxicity and opinion congruence. Participants were more willing to engage with a civil discourse thread even when their opinion was in the minority than they were to engage with a toxic thread even when their opinion was in the majority (Cohen's d = 0.25, SE = 0.11, $p < 0.05$). In other words, our findings show that the fear of engaging in a toxic thread outweighs the fear of expressing a minority view in a civil context, offering a nuanced extension to the spiral of silence theory. In addition, we did not observe any effect between the topics in their respective civil and toxic discourse groups.

Significant predictors for willingness to engage included a participant's history of posting on social media (Cohen's d = 0.81, SE = 0.24, $p < 0.01$) and their self-reported knowledge on the topic (Cohen's d = 0.77, SE = 0.15, $p < 0.01$).

4.2 Qualitative Findings

We present our qualitative findings on the influence of AI summary tools, organised into five themes reflecting participants' opinions.

4.2.1 *Positive Views Towards the AI Summary Tools.*

Perceived Usefulness and Effectiveness: Participants perceived the AI summary tools as effective and useful for understanding the comment section of the Reddit threads due to its ability to identify key talking points to grasp the (perceived) majority view of the online public (thus seeing social media as a window for viewing what society thinks). Forty participants provided a positive response towards the AI summarisation tool, as they thought the tool *"worked well"* ($P7_{Both}$) or *"worked very well"* ($P11_{Narratives}$), and eighteen participants expressed that the tool was useful, with claims that: *"It was helpful to see the summarised information first before filtering through the conversation comment by comment. It was kind of like having bullet points to focus on beforehand."* ($P105_{Narratives}$). As conveyed by the aforementioned quote, thirty-six participants specifically found the summarisation of the comment section helpful in providing an overview for both sides of the argument. As stated by $P39_{Narratives}$: *"I felt that AI had a great deal on helping me decide on whether I support the topic or not. It provided clear information on both sides."* Moreover, five participants perceived benefits of being presented arguments from both sides in a *"neutral tone"* ($P82_{Both}$), particularly for *"emotionally charged issues"* ($P24_{Both}$).

Additionally, fifteen participants found that the tool effectively *"condenses complex topics"* ($P53_{Narratives}$ & $P82_{Both}$) by *"providing clarity"* ($P70_{Narratives}$), which made it easier for them to understand the topic of the posts. Subsequently, six participants stated that this tool saved them time from reading the threads: *"The tool made it very easy to grasp the main points being made in the thread. Saves you the time you would use to go through each comment."* ($P67_{Narratives}$).

4.2.2 *Limitations and Risks of AI Summary Tools.*

Concerns toward AI Tool Usage: The main concerns expressed by participants towards the AI summaries was whether they felt it was needed in an era where AI and a lack of human oversight is becoming more prevalent in social media. Three participants disliked the tool as they felt they *"do not need it"* ($P87_{Percentages}$), and three

participants found the summarisation tool to be inaccurate. Four participants expressed concern over the usage of these tools as it could potentially be influenced by bots, in addition to the perceived drawback that the summaries are *"people's opinions and not facts"* ($P71_{Narratives}$). One rare subtheme identified by two participants was the conformity risk of these AI summaries: *"It was a nice tool that saves the user from having to read every comment to see the "vibe" of the comment section. Right after the title and picture you get a statistic from the ai tool that gives you the general consensus of the comments. It was nice, but also can provide initial bias to people who are not informed on the topic. If a topic is 90% negative and 10% correct perhaps the viewer will see the statistic and lean towards the majority side if they are a close minded person and follow the pack mentality."* ($P129_{Percentages}$).

Reservations Related to the Summarisation of Comments: Participants appreciated the tool's perceived neutrality and utility as a proxy to the comments, though at the risk of losing conversational nuance. Seven participants voiced doubt in the summarisation of the comment thread as nuance or the tone of individual comments could have been overlooked: *"The AI summarisation tool provides a clear and concise overview of complex topics, which is helpful for quickly understanding multiple perspectives. It simplifies the process of digesting large amounts of information and makes it easier to engage with diverse viewpoints. However, I believe it could benefit from including more nuanced details or context, as sometimes the summaries feel too brief to fully capture the depth of certain arguments."* ($P110_{Both}$). Subsequently, this could potentially cause further discussion comments to be superficial: *"The tool is effective for providing concise, focused insights on complex topics. It does a good job of distilling key points while maintaining a neutral tone. It's helpful in situations where quick understanding is needed, but the summaries might lack some depth for deeper discussions."* ($P91_{Percentages}$). Two participants expressed concerns over such tools helping platform fringe issues, highlighting that, *"it can be a helpful way to see reasons on both sides of an argument, but some 'arguments' shouldn't really have two sides."* ($P1_{Narratives}$).

4.2.3 *Influence of AI Summary Tools on Personal Judgement.*

Interpretation of Majority Consensus: The AI tools acted as a proxy for identifying the perceived 'ground truth' of the majority opinion, with participants relying more on the AI than on their own analysis of the polarity of the comment section. Six participants reported that the AI summarisation tool helped them interpret the *"dominant perspective"* ($P116_{Narratives}$) that subsequently allowed them to *"spot which comments aligned with which perspective"* ($P24_{Both}$). Participants found this particularly beneficial for being aware of the overall sentiment prior to engaging with the thread: *"It was pretty helpful! I liked that I was able to see a broad view of the tenor of the conversation beforehand."* ($P56_{Percentages}$). Additionally, five participants further explained that the tool *"helped support opinions and discussions"* ($P10_{Narratives}$), thus expediting the formation of their personal opinions: *"The summaries captured the key points and tone of the discussions fairly well, which helped me form or reflect on my opinions faster."* ($P28_{Percentages}$).

Perceived Objectivity: Overall, participants found the tool as objective and balanced, despite often incorrectly estimating the level of polarisation in the AI summaries only condition. Eighty-nine participants found the AI summarisation tool to be "*fair*" or "*balanced*". Some found that the neutral tone was *"..especially valuable for contentious subjects like the Russia-Ukraine ceasefire or semi-automatic rifle bans."* ($P82_{Both}$). However, one participant pointed out a preference for the tool to reflect the tone of the threads: *"It didn't characterise the tone of the discussions in the threads. The Musk thread was mostly insults. Very aggressive. The Fukushima discussion was intelligent and filled with healthy discourse. The AI treated both equally. Not great."* ($P116_{Narratives}$).

5 Discussion

In this paper, we show that individuals exhibit more substantial opinion shifts when exposed to AI summarisation tools. As such, our findings offer a double-edged sword for HCI research. On one hand, our qualitative findings identified the tool as a helpful aide to quickly identify common narratives, help improve reading comprehension, and contribute to their opinion-making process. On the other hand, this very process risks reinforcing echo chambers and group mentality by entrenching division (via visualising the percentages), or oversimplifying nuanced topics through summaries as presently seen on Facebook [60], 𝕏 [68], and Reddit [91].

Thus, in this section, we outline the implications of the tools in enhancing social conformity and its role in social media as a result of our study, the challenge of detoxifying conversations and its ethical implications, as well as the overall implications for HCI researchers and platforms themselves. We then conclude with the limitations and future work in the field of AI-assisted tools for social media platforms.

5.1 The Role of Individualism and Group Dynamics in the Use of AI Summary Tools

Our findings reveal a contrasting divide: while most participants subjectively experienced the tools as aids for individual opinion-making (perceiving them as helpful and neutral), our quantitative results show that they really function as instruments of group coercion. The presence of both AI tools increased the percentage of participants conforming to the majority opinion from 34% to 42%. To deconstruct this divide, we relate back to our research questions:

RQ1: How do AI-generated summaries of ideological perspectives in social media threads influence users' opinions?

Making the majority opinion explicit using the AI percentages tool significantly intensified social conformity compared to the no-AI control group, even though both groups had the ability to 'check for themselves' and identify the level of polarity in the comments. This aligns with *Social Identity Theory* [33], since the quantified percentage acted as a form of heuristic *social proof* (joining in for what is seen as 'right' by the majority, as applicable in the polarised threads) that made the in-group more visible and salient, which triggered a desire for social validation through normative influence. The fact that our percentages and narratives tool had any effects on user opinion is concerning given the rapid development and deployment of AI in social media with zero regard as to how it could impact our democratic discourse and opinions. Designing AI for social media comes with great responsibility for its noteworthy ability to influence electoral outcomes and offline polarisation [8,

25, 57, 97]. Platforms could incorporate informational summaries that include both factuality assessments and percentages of political leaning, thereby promoting group perceptions that are grounded in accuracy and awareness of ideological bias. However, social media platforms currently only offer informational summaries and warnings on cases of false information (mis/disinformation), unlike narrative summaries in all comment sections [60]. Thus, platforms should disclose and be cognisant of the potential biases that these models can have on public opinion, as discussed in this research, rather than engage in the critiqued techno-solutionism of placing AI in all aspects of social life [12].

Furthermore, while AI narrative summaries did not depolarise participants, they did moderate the *magnitude* of opinion change by creating a false balance effect, similar to the issue of climate denialism found online due to the disproportionate information online [13, 96].

RQ2: To what extent do AI-generated summaries of ideological perspectives impact users' likelihood of engaging with a social media thread?

The AI tools did not significantly affect participants' desire to engage/respond to the thread; instead, overall thread toxicity was the most powerful predictor. This challenges the primacy of the *Spiral of Silence* theory [64], where the anticipated emotional cost of engaging with toxic discourse on a polarising thread topic outweighed the benefit of participation. Interestingly, platforms that require real names are more prone to the spiral of silence effect due to the potential for offline/real-world consequences [34, 88, 90], while username-based pseudo-anonymous platforms typically cause the *opposite effect*, that people become emboldened to share their opposing views. This contradicting effect is known as the *online disinhibition effect* [99], referring to how online anonymity can embolden otherwise hidden views when communicating online compared to in-person or identity-attributable platforms. One potential explanation for the lack of participants' desire to engage in our pseudo-anonymous subreddits could be due to the nature of the questions relating to politics and the documented mental burn-out observed after the 2024 election leading to disengagement from these topics [1]. Overall, while we did not observe any significant difference within each topic in their respective civil/toxic categories, the effect of toxic discourse in a thread stifled their willingness to engage compared to the civil discourse threads.

In addition, our qualitative data highlighted that participants appreciated the 'neutral tone' of summaries for making the topics feel more approachable, while one noted that the AI's failure to capture a thread's 'aggressive' tone was a flaw. This implies that a summary's 'self-censorship' can end up sanitising the conversational tone but can foster engagement.

This finding presents a core design dilemma for platforms: if a summary sanitises a thread with toxic discourse, it provides a false belief of a productive discussion, which can erode user trust when they scroll through and encounter hostility. Conversely, accurately reflecting toxicity may reinforce a negative feedback loop, where we observed that by showing users the high-polarity of a thread through the AI percentages tool, it made them more likely to be polarised on the thread, thus making it challenging to break free from polarisation. Navigating this trade-off between representational accuracy and user engagement is a central ethical challenge. This may point to future work in adaptive interfaces, where users could observe factual informational summaries first from an AI tool, before setting their preference for the summary of the conversations (e.g., "Show me the polite version" vs. "Show me the raw discussion"), allowing for more user agency in navigating online discourse.

5.1.1 Balancing Informational and Social Conformity Push and Pulls.

To mitigate the pull of group consensus, the future objective of AI summarisation tools should consider encouraging critical opinion-making at multiple stages of user interaction: before a user opens a thread, while they are reading *informational* sources (news sources, statement-of-facts etc.), and as they engage with others' *opinions*. The current blunt approach of summarising entire conversations, as seen by Meta [60], 𝕏 [68], and Reddit [91], only offers social but not necessarily informational conformity. A more nuanced approach would involve deploying different types of summaries and percentages that help users distinguish between verifiable information and social opinion. For instance, platforms could provide summaries of factual claims (with integrated fact-checking), alongside analytics on the sources being cited (e.g., the percentage of *news sources* covering the topic and their political leaning). This would separate the analysis of information from the analysis of public opinion.

Research into prebunking/inoculation of conformity remains sparse for AI summarisation tools. Nonetheless, Roozenbeek and van der Linden identified that participants who played an online game designed to teach common misinformation tactics (e.g., emotional language, polarisation) were later more resistant to misleading comments and content [86]. Likewise, community notes offer an example of an *informational source summary*, which appear below the post but above the comments [111]. As opposed to interactive experiences, Ecker et al. identified that explicit warnings and notes had a weak to moderate effect at reducing the spread of polarised threads containing misinformation [21]. Thus, social media platforms must consider implementing multiple safeguards while not eliminating the ease and purpose of implementing AI—to summarise and *not* over-complicate things. After all, participants perceived the tool as most useful and effective to determine whether to read, or how much effort to put into reading, an online thread—as they could ascertain the content of the discussion in the condensed summary. However, beyond embedding safeguards, participants highlighted the need for nuance and tone/irony detection.

AI summarisation tools can also represent a potential vector for manipulation as our findings display their ability to persuade and shape opinions, which Fogg et al. frames as 'captology' technologies [22]. For example, a political campaign could microtarget users with threads whose AI summaries are curated to show a manufactured consensus, relying on a 'both sides' approach to platform and mainstream a fringe political view. This could inadvertently normalise and platform fringe beliefs, which could make them more palpable by the population due to its false perceived popularity. Our results highlight this risk in our AI narratives only condition, where participants misconstrued the real proportion of commenters' opinions despite having the time and ability to read the comments themselves as they overrelied on the AI narratives tool (Figure 9). This presents a significant design challenge for platforms: how can

summaries provide 'at-a-glance' overviews of conversations that users desire without this data becoming a tool for conformity? If they provide both contexts of the proportion of commenters' beliefs *with* the narrative summaries, then this reinforces the conformity towards the majority group—but one can ask if this is any better? If used maliciously, such a feature could be considered a 'dark pattern' in social media's UX design [29], deceptively steering user opinion. This points toward a need for research into counter-persuasive interfaces that could encourage users to read a random selection of comments first before showing the group consensus.

5.2 AI's Role in Detoxifying and Reframing Contentious Conversations

Beyond conformity, one of the main concerns of AI summaries is whether to platform 'both sides'. Embodying this claim was the following participant statement: *"both talking points would be given equal weight, leading a person to find inaccuracies treated with legitimacy. That's dangerous. Racist messages and outright lies are given validity with this AI tool."* Thus, AI summaries platform all available perspectives, including those that may contain misinformation or those that are extreme/outside of society's accepted range of political beliefs. Social media platforms should take note of these findings lest AI summaries inadvertently alter users' perceptions of the truth (without robust fact-checking integration) or consensus (which could platform harmful content). Nonetheless, the issue of platforming minority or new opinions vs. what constitutes 'harmful' perspectives is a contentious area of fact-checking and algorithmic-design research [28, 38, 39, 55]. Furthermore, our participants identified that the neutral, analytical and objective framing of the narrative summaries made the tool seem more trustworthy and fair, particularly in the toxic threads, at the expense of losing the ability to capture the thread's animosity and tone. As such, analysing and summarising online argumentation helps users prepare counter-arguments by understanding the main points in a conversation without the unnecessary insults/toxicity, similar to related work in using AI to summarise debates to help students prepare informative rebuttals [49], or using AI as a tool to help promote critical thinking by detoxifying comments and identifying key areas of disagreement to resolve [28, 75].

Thus, the balance between addressing toxicity, misinformation and whether to platform certain ideas should be the role of an independent tool to avoid delegitimising the authenticity and utility of an objective AI summariser. After all, there is little utility of a biased AI summariser as its role is to simply condense and summarise the sentiment and opinions of the debaters.

Nonetheless, some participants doubted the ability of any AI to remain impartial when summarising information—raising concerns on what talking points and information to prioritise in a condensed summary which may lack nuance. This scepticism points to a deeper issue: the impossibility of *true* neutrality in summarisation. Every decision about what to include or exclude, how to frame arguments, and what language to use inherently involves value judgements that reflect the biases of both the AI system, even without specific developer intervention.

Future work should explore balancing nuance with community involvement in AI narrative summaries, potentially through a Delphi method for automated summaries—starting with a baseline summary, allowing anonymous community engagement, and polling users of different persuasions to verify the veracity of summary variants to create a more reflective summary that each side can agree to. Thus, summaries would be decentralised and deliberative, rather than risking opaque or even malicious summaries by vested parties. This is particularly notable with the rise of preprogrammed responses in LLMs as seen in the global geopolitics of LLM development [42, 65, 119]. Furthermore, summaries could delineate between verified fact-checked information summaries vs. opinion-based 'sides', both with the human-in-the-loop community-based verification. Exploring these approaches, before deploying them live, would further ensure that social AI tools become responsible AI tools and help protect against inadvertent consensus-driven polarisation.

5.3 Implications for HCI Research and Practice

Our results fundamentally demonstrate that AI summary tools can enhance the persuasive effect of social conformity from comment sections. This is of particular concern as social media platforms deploy summariser tools in the hopes that it helps reading comprehension and the analysis of long and potentially toxic discussion threads, but neither companies nor researchers have considered the role these tools can have in exacerbating social conformity or polarisation. Moving forward, industry and HCI researchers must consider the psychological consequences of AI tools *before* deploying them publicly, particularly due to the risks to democratic integrity that saturating social media with AI could have in reshaping our political perceptions. Highlighting this potential for abuse, 𝕏's integration of the Grok chatbot included a "programming error" that led to it providing summaries platforming holocaust denial [42], and propelling topics such as the alleged 'white genocide' in South Africa on unrelated topics [44], with the latter being claimed by xAI as the result of an unauthorised modification by a rogue employee [17, 84, 112]. Resolving these issues requires explainability and attribution for summary-bot designs, such as transparent open-sourced prompts to display how the bot produces its summaries, as well as UX design considerations—such as providing informational summaries to enable users to form an early opinion before reading a separate delineated summary of social *opinions*.

Given the conformity effect we found on AI-generated percentages, developers should consider time delays or an option to explicitly click-to-request a summary or percentage breakdown to encourage users to think critically before emotively or impulsively reacting to the group mentality promoted by these AI tools. Similarly, Masrani et al. found that delaying additional posts in a heated debate reduced opinion polarity, toxicity, and promoted healthy on-topic discussions [59]; where future work should investigate if this conversational trend transfers when analysing comment sections.

5.4 Limitations and Future Work

Our study simulates online discussions where users hold strong one-sided pro- or anti-topic opinions across six threads with either toxic or civil discourse. In doing so, we follow related work in simulating

online conversations to control for the potential confounds such as the commenters' academic knowledge, English-language competency, and post length, to ensure that each pro and anti topic comment had approximately the same conditions/weight as each other, utilising the current state-of-the-art GPT-4.5 language model [69]. Moreover, our simulation approach addresses the ethical issue of in-the-wild social media experiments, given the backlash of a 2025 study in which researchers deployed synthetic AI tools and posts on Reddit without obtaining consent, ultimately resulting in both reputational and ethical repercussions [98]. Overall, our study provides the foundation for how AI summarisation tools can reinforce social conformity in a clear and controlled manner. Future work should also consider the degrees of agreement (such as percentages of strong agreement vs. neutral or mild dis/agreement) to test the role of nuance and changing opinions (from the commenters themselves) as a factor of future AI summarisation design.

Likewise, simulating pro- and anti- topic GPT-4.5 prompts ensures that we can generate and verify that each post reflects the correct stance and to ensure the percentages we use for the 'AI' tool are correct. However, AI summariser tools are still limited based on the performance of the language model, which can still have issues understanding uncertainty [69], cultural nuance [108] or layers of meta-irony often found to evade AI-driven content-moderation systems or for memetic culture (e.g., taking an extreme stance deliberately to highlight its obscenity) [27, 28, 93]. Thus, future work should consider the potential damaging consequences of how *incorrect* assessments of threads, such as classifying a thread as majority for a topic when it is actually against, can have in the battle for social conformity. This could open interesting findings, as participants in our study had the ability/time to read the summaries and 'verify' themselves by reading the comments. If there were to be a disconnect between these AI tools and the real ground truth comment section, this could further highlight our societal risk of over-reliance on AI systems, and undermine our social and democratic cohesion.

Our study highlights salient examples of threads with toxic or civil discourse to balance the coverage of topics (including local and international news) and destructive/constructive discourse in threads where they are most likely to occur. While this design balances topical diversity and time (to prevent reading fatigue, which could undermine our results), future work should consider threads that lie in-between civil threads and the higher-stakes 'fate-of-the-nation' threads with toxic discourse (e.g., Russia–Ukraine war). Focusing on variations of discourse civility/toxicity across topics would further our understanding of the dynamics of AI tools on shaping user opinion.

In addition to testing cases where opinions may be split, the effect of confounds such as karma and style of replies should also be observed to test how AI tools shape perceptions across different formats and platforms. For instance, while our focus is on the Reddit-style 'post, comment-reply' format, other threads or platforms may foster deeper or shallower-style conversations. While Reddit typically has 1–2 levels of reply depth as replicated [23, 109], debating platforms such as Kialo target long 1 vs. 1 debate chains [45], while real-time streaming platforms like Twitch target quick shallow-depth comments without sub-replies. Thus, for studies that go beyond *social media comment sections* and into the streaming space, future work could draw inspiration from the real-time opinion visualisation tool such as 'The Worm'—a tool for gauging an audience's opinion and displaying it in real-time as previously used in televised debates in the UK [18], US [19], Australia [3], and New Zealand [30]. However, critics oppose the use of these real-time individualised opinion aggregation tools due to the risk of undermining critical thought and reinforcing groupthink, delegitimising the role of informational conformity as reflected in our study [19].

6 Conclusion

As we enter a world where AI appears everywhere, the risks to our critical thinking and individualism come under threat. Whether AI in our media enables a potential dystopian groupthink depends on whether researchers and industry investigate its potential risks. Our investigation on the role of these social media tools, both summarising percentages of agreement and narrative summaries, highlights the role they have in reinforcing social conformity and enhancing polarisation (for the *AI percentages* tool), as well as the potential to hinder/reduce opinion changes (for the *AI narratives* tool). As these tools shape our opinions and perception of polarisation online, developers should consider mitigations that occur before, during, and after a user engages with a social media thread. Our findings highlight the need to develop AI tools that accentuate factual *informational* conformity and critical thought before users view comment summaries, with the objective to help the user create their own opinion independently based on factual evidence, rather than abiding by the conformitive normality of the social media mob mentality.

References

[1] AP-NORC. 2024. *Most adults feel the need to limit political news consumption due to fatigue and information overload.* Associated Press-NORC Center for Public Affairs Research. https://apnorc.org/projects/most-adults-feel-the-need-to-limit-political-news-consumption-due-to-fatigue-and-information-overload/

[2] Pablo Aragón, Vicenç Gómez, David García, and Andreas Kaltenbrunner. 2017. Generative models of online discussion threads: state of the art and research challenges. *Journal of Internet Services and Applications* 8, 1 (2017), 15. doi:10.1186/s13174-017-0066-z

[3] Lincoln Archer and Clinton Poreous. 2007. *Rudd given nod in close debate.* Courier Mail. https://web.archive.org/web/20071023031645/http://www.news.com.au/couriermail/story/0,23739,22624834-952,00.html/

[4] S. E. Asch. 1951. *Effects of group pressure upon the modification and distortion of judgments.* Carnegie Press, Oxford, England, 177–190.

[5] Cristina Aybar, Virgilio Pérez, and Jose M. Pavía. 2024. Scale matters: unravelling the impact of Likert scales on political self-placement. *Quality & Quantity* 58, 4 (2024), 3725–3746. doi:10.1007/s11135-023-01825-2

[6] Stefano Balietti, Lise Getoor, Daniel G. Goldstein, and Duncan J. Watts. 2021. Reducing opinion polarization: Effects of exposure to similar people with differing political views. *Proceedings of the National Academy of Sciences* 118, 52 (2021), e2112552118. doi:10.1073/pnas.2112552118

[7] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit Dataset. arXiv:2001.08435 [cs.SI] https://arxiv.org/abs/2001.08435

[8] Alonso Bernal, Cameron Carter, Ishpreet Singh, Kathy Cao, and Olivia Madreperla. 2020. *Cognitive Warfare: An Attack on Truth and Thought.* NATO and Johns Hopkins University. https://www.innovationhub-act.org/sites/default/files/2021-03/Cognitive%20Warfare.pdf

[9] Md Momen Bhuiyan, Michael Horning, Sang Won Lee, and Tanushree Mitra. 2021. NudgeCred: Supporting News Credibility Assessment on Social Media Through Nudges. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 427 (Oct. 2021), 30 pages. doi:10.1145/3479571

[10] Md Momen Bhuiyan, Sang Won Lee, Nitesh Goyal, and Tanushree Mitra. 2023. NewsComp: Facilitating Diverse News Reading through Comparative Annotation. 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 548, 17 pages. doi:10.1145/3544548.3581244

[11] Levi Boxell, Matthew Gentzkow, and Jesse M. Shapiro. 2024. Cross-Country Trends in Affective Polarization. *The Review of Economics and Statistics* 106, 2 (2024), 557–565. doi:10.1162/rest_a_01160

[12] Danah Boyd. 2025. *We Need an Interventionist Mindset.* Tech Policy Press. https://www.techpolicy.press/we-need-an-interventionist-mindset/

[13] Maxwell T Boykoff and Jules M Boykoff. 2004. Balance as bias: global warming and the US prestige press. *Global Environmental Change* 14, 2 (2004), 125–136. doi:10.1016/j.gloenvcha.2003.10.001

[14] Curtis Bram. 2024. Beyond partisan filters: Can underreported news reduce issue polarization? *PLOS ONE* 19, 2 (2024), 1–11. doi:10.1371/journal.pone.0297808

[15] Megan Brenan. 2022. *Americans' Trust In Media Remains Near Record Low.* Gallup Inc. https://news.gallup.com/poll/403166/americans-trust-media-remains-near-record-low.aspx

[16] Jacob Cohen. 1992. A power primer. *Psychological Bulletin* 112, 1 (1992), 155–159. doi:10.1037/0033-2909.112.1.155

[17] Kate Conger. 2025. *Employee's Change Caused xAI's Chatbot to Veer Into South African Politics.* The New York Times. https://www.nytimes.com/2025/05/16/technology/xai-elon-musk-south-africa.html

[18] Colin Davis. 2014. *Evidence on Broadcast General Election Debates.* 55th Parliament of the United Kingdom. https://committees.parliament.uk/writtenevidence/47595/html/

[19] Colin J. Davis, Jeffrey S. Bowers, and Amina Memon. 2011. Social Influence in Televised Election Debates: A Potential Distortion of Democracy. *PLOS ONE* 6, 3 (03 2011), 1–7. doi:10.1371/journal.pone.0018154

[20] Elizabeth Dubois and Julia Szwarc. 2018. Self-Censorship, Polarization, and the—Spiral of Silence on Social Media. In *Policy & Politics Conference.* Oxford Internet Institute, Oxford, UK.

[21] Ullrich K. H. Ecker, Stephan Lewandowsky, and David T. W. Tang. 2010. Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition* 38, 8 (2010), 1087–1100. doi:10.3758/MC.38.8.1087

[22] B. J. Fogg. 1999. Persuasive technologies. *Commun. ACM* 42, 5 (May 1999), 26–29. doi:10.1145/301353.301396

[23] Maria Glenski, Corey Pennycuff, and Tim Weninger. 2017. Consumers and Curators: Browsing and Voting Patterns on Reddit. *IEEE Transactions on Computational Social Systems* 4, 4 (2017), 196–206. doi:10.1109/TCSS.2017.2742242

[24] Drew Gorenz and Norbert Schwarz. 2024. How funny is ChatGPT? A comparison of human- and A.I.-produced jokes. *PLOS ONE* 19, 7 (07 2024), 1–13. doi:10.1371/journal.pone.0305364

[25] 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. doi:10.1145/3583067

[26] Jarod Govers, Philip Feldman, Aaron Dant, and Panos Patros. 2023. Prompt-GAN–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. doi:10.1145/3594315.3594366

[27] Jarod Govers, Saumya Pareek, Eduardo Velloso, and Jorge Goncalves. 2025. Feeds of Distrust: Investigating How AI-Powered News Chatbots Shape User Trust and Perceptions. *ACM Trans. Interact. Intell. Syst.* 15, 4, Article 20 (Dec. 2025), 31 pages. doi:10.1145/3722227

[28] 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* (Honolulu, HI, USA) *(CHI '24).* Association for Computing Machinery, New York, NY, USA, 18 pages. doi:10.1145/3613904.3642322

[29] Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs. 2018. The Dark (Patterns) Side of UX Design. 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, 1–14. doi:10.1145/3173574.3174108

[30] Duncan Greive. 2003. *The worm electrocuted politics in 2002. Now we're living in the worm's world.* The Spinoff. https://thespinoff.co.nz/politics/12-10-2023/the-worm-electrocuted-politics-in-2002-now-were-living-in-the-worms-world/

[31] Ground News. 2024. *Methodology - Media Bias Rating System.* Snapwise Inc. https://ground.news/rating-system#biasRating

[32] Keith N Hampton, Harrison Rainie, Weixu Lu, Maria Dwyer, Inyoung Shin, and Kristen Purcell. 2014. Social media and the'spiral of silence'.

[33] Tajfel Henri and John C. Turner. 1979. . Brooks/Cole, Monterey, CA, Chapter An integrative theory of intergroup conflict, 33–47.

[34] Christian Pieter Hoffmann and Christoph Lutz. 2017. Spiral of Silence 2.0: Political Self-Censorship among Young Facebook Users. In *Proceedings of the 8th International Conference on Social Media & Society* (Toronto, ON, Canada). Association for Computing Machinery, New York, NY, USA, Article 10, 12 pages. doi:10.1145/3097286.3097296

[35] Myiah J. Hutchens, David E. Silva, Jay D. Hmielowski, and Vincent J. Cicchirillo and. 2019. What's in a username? Civility, group identification, and norms. *Journal of Information Technology & Politics* 16, 3 (2019), 203–218. doi:10.1080/19331681.2019.1633983

[36] International Atomic Energy Agency. 2023. *IAEA Finds Japan's Plans to Release Treated Water into the Sea at Fukushima Consistent with International Safety Standards.* https://www.iaea.org/newscenter/pressreleases/iaea-finds-japans-plans-to-release-treated-water-into-the-sea-at-fukushima-consistent-with-international-safety-standards

[37] Farnaz Jahanbakhsh and David R Karger. 2024. A Browser Extension for in-place Signaling and Assessment of Misinformation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24).* Association for Computing Machinery, New York, NY, USA, Article 946, 21 pages. doi:10.1145/3613904.3642473

[38] Farnaz Jahanbakhsh, Yannis Katsis, Dakuo Wang, Lucian Popa, and Michael Muller. 2023. Exploring the Use of Personalized AI for Identifying Misinformation on Social Media. 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 105, 27 pages. doi:10.1145/3544548.3581219

[39] Farnaz Jahanbakhsh, Amy X. Zhang, and David R. Karger. 2022. Leveraging Structured Trusted-Peer Assessments to Combat Misinformation. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 524 (nov 2022), 40 pages. doi:10.1145/3555637

[40] Kokil Jaidka, Subhayan Mukerjee, and Yphtach Lelkes. 2023. Silenced on social media: the gatekeeping functions of shadowbans in the American Twitterverse. *Journal of Communication* 73, 2 (01 2023), 163–178. arXiv:https://academic.oup.com/joc/article-pdf/73/2/163/49678253/jqac050.pdf doi:10.1093/joc/jqac050

[41] Jonas L. Juul and Johan Ugander. 2021. Comparing information diffusion mechanisms by matching on cascade size. *Proceedings of the National Academy of Sciences* 118, 46 (2021), e2100786118. arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2100786118 doi:10.1073/pnas.2100786118

[42] Ashifa Kassam. 2025. *Musk's AI bot Grok blames 'programming error' for its Holocaust denial.* The Guardian. https://www.theguardian.com/technology/2025/may/18/musks-ai-bot-grok-blames-its-holocaust-scepticism-on-programming-error

[43] Scott Keeter. 2022. *Public Opinion Polling Basics.* Pew Research Center. https://www.nbcnewyork.com/new-york-city/andrew-cuomo-mayor-bus-plan/6177890/

[44] Data Kerr. 2025. *Musk's AI Grok bot rants about 'white genocide' in South Africa in unrelated chats.* The Guardian. https://www.theguardian.com/technology/2025/may/14/elon-musk-grok-white-genocide

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

[46] Ioannis Kontostathis, Evlampios Apostolidis, Konstantinos Apostolidis, and Vasileios Mezaris. 2025. Enhancing User Control in AI-Based Video Summarization for Social Media. In *MultiMedia Modeling: 31st International Conference on Multimedia Modeling, MMM 2025, Nara, Japan, January 8–10, 2025, Proceedings, Part V* (Nara, Japan). Springer-Verlag, Berlin, Heidelberg, 119–126. doi:10.1007/978-981-96-2074-6_12

[47] Robert Kubinec. 2025. *Introduction to ordbetareg.* https://cran.r-project.org/web/packages/ordbetareg/vignettes/package_introduction.html

[48] Samya Kullab. 2025. *Things to know about the limited ceasefire between Russia and Ukraine brokered by the US.* Associated Press. https://apnews.com/article/ukraine-russia-us-limited-ceasefire-4f1d4a835c52e8a37716ea21b32ccb0b

[49] Elmar Kutsch. 2023. *Harness human and artificial intelligence to improve classroom debates.* Times Higher Education. https://www.timeshighereducation.com/campus/harness-human-and-artificial-intelligence-improve-classroom-debates

[50] Issie Lapowsky. 2018. *NewsGuard Wants to Fight Fake News With Humans, Not Algorithms.* Wired. https://www.wired.com/story/newsguard-extension-fake-news-trust-score/

[51] Z. Leviston, I. Walker, and S. Morwinski. 2013. Your opinion on climate change might not be as common as you think. *Nature Climate Change* 3, 4 (2013), 334–337. doi:10.1038/nclimate1743

[52] Rebecca Lewis. 2018. Alternative influence: broadcasting the reactionary right on YouTube. https://datasociety.net/wp-content/uploads/2018/09/DS_Alternative_Influence.pdf

[53] Mengqi Liao, S. Shyam Sundar, and Joseph B. Walther. 2022. User Trust in Recommendation Systems: A comparison of Content-Based, Collaborative and Demographic Filtering. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22).* Association for Computing Machinery, New York, NY, USA, Article 486, 14 pages. doi:10.1145/3491102.3501936

[54] Ziyue Lin, Yi Shan, Lin Gao, Xinghua Jia, and Siming Chen. 2025. SimSpark: Interactive Simulation of Social Media Behaviors. *Proc. ACM Hum.-Comput.*

Interact. 9, 2, Article CSCW168 (May 2025), 32 pages. doi:10.1145/3711066

[55] Zhuoran Lu, Patrick Li, Weilong Wang, and Ming Yin. 2022. The Effects of AI-based Credibility Indicators on the Detection and Spread of Misinformation under Social Influence. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 461 (Nov 2022), 27 pages. doi:10.1145/3555562

[56] Robert Luzsa and Susanne Mayr. 2021. False consensus in the echo chamber: Exposure to favorably biased social media news feeds leads to increased perception of public support for own opinions. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 15, 1 (Feb. 2021), Article 3. doi:10.5817/CP2021-1-3

[57] Megan MacDuffee Metzger and Joshua A. Tucker. 2017. Social Media and EuroMaidan: A Review Essay. *Slavic Review* 76, 1 (2017), 169–191. doi:10.1017/slr.2017.16

[58] Maruti Techlabs. 2016. *News Bots are Changing The Way we Read News.* Medium. https://chatbotsmagazine.com/news-made-personal-with-chatbots-6dbba0691475

[59] 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. doi:10.1145/3579523

[60] Meta. 2023. *Building Generative AI Features Responsibly.* https://about.fb.com/news/2023/09/building-generative-ai-features-responsibly/

[61] Matto Mildenberger and Dustin Tingley. 2019. Beliefs about Climate Beliefs: The Importance of Second-Order Opinions for Climate Politics. *British Journal of Political Science* 49, 4 (2019), 1279–1307. doi:10.1017/S0007123417000321

[62] Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2024. More human than human: measuring ChatGPT political bias. *Public Choice* 198, 1 (2024), 3–23. doi:10.1007/s11127-023-01097-2

[63] Jimin Mun, Cathy Buerger, Jenny T Liang, Joshua Garland, and Maarten Sap. 2024. Counterspeakers' Perspectives: Unveiling Barriers and AI Needs in the Fight against Online Hate. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 742, 22 pages. doi:10.1145/3613904.3642025

[64] Elisabeth Noelle-Neumann. 1974. The Spiral of Silence A Theory of Public Opinion. *Journal of Communication* 24, 2 (1974), 43–51. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1460-2466.1974.tb00367.x doi:10.1111/j.1460-2466.1974.tb00367.x

[65] Sander Noels, Guillaume Bied, Maarten Buyl, Alexander Rogiers, Yousra Fettach, Jefrey Lijffijt, and Tijl De Bie. 2025. What Large Language Models Do Not Talk About: An Empirical Study of Moderation and Censorship Practices. arXiv:2504.03803 [cs.CL] https://arxiv.org/abs/2504.03803

[66] Ruchi Ookalkar, Kolli Vishal Reddy, and Eric Gilbert. 2019. Pop: Bursting News Filter Bubbles on Twitter Through Diverse Exposure. In *Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing* (Austin, TX, USA) *(CSCW '19 Companion)*. Association for Computing Machinery, New York, NY, USA, 18–22. doi:10.1145/3311957.3359513

[67] OpenAI. 2024. *A landmark multi-year global partnership with News Corp.* https://openai.com/index/news-corp-and-openai-sign-landmark-multi-year-global-partnership/

[68] OpenAI. 2025. *About Grok, Your Humorous AI Assistant on X.* https://help.x.com/en/using-x/about-grok/

[69] OpenAI. 2025. *OpenAI GPT-4.5 System Card.* https://cdn.openai.com/gpt-4-5-system-card-2272025.pdf

[70] Saumya Pareek and Jorge Goncalves. 2024. Peer–supplied credibility labels as an online misinformation intervention. *International Journal of Human-Computer Studies* (2024), 41 pages. doi:10.1016/j.ijhcs.2024.103276

[71] M. C. Parent, T. D. Gobble, and A. Rochlen. 2019. Social Media Behavior, Toxic Masculinity, and Depression. *Psychol Men Masc* 20, 3 (2019), 277–287. doi:10.1037/men0000156

[72] Eli Pariser. 2011. *The Filter Bubble: What the Internet Is Hiding from You.* The Penguin Group, New York, NY.

[73] Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. 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 74, 18 pages. doi:10.1145/3526113.3545616

[74] Adam Piore. 2018. *Technologists are trying to fix the "filter bubble" problem that tech helped create.* MIT Technology Review. https://www.technologyreview.com/2018/08/22/2167/technologists-are-trying-to-fix-the-filter-bubble-problem-that-tech-helped-create/

[75] Priya Pitre and Kurt Luther. 2024. ArguMentor: Augmenting User Experiences with Counter-Perspectives. arXiv:2406.02795 [cs.HC] https://arxiv.org/abs/2406.02795

[76] John O. Rawlings, Sastry G. Pantula, and David A. Dickey. 1998. *Class Variables in Regression* (2nd ed.). Springer New York, New York, NY, 269–323. doi:10.1007/0-387-22753-9_9

[77] Reddit. 2025. *Archive r/moderatepolitics (Bill banning social media for youngsters advances) (March 2025).* https://web.archive.org/web/20250326025330/https://www.reddit.com/r/moderatepolitics/comments/1iil327/bill_banning_social_media_for_youngsters_advances/

[78] Reddit. 2025. *Archived r/politics (March 2025).* https://web.archive.org/web/20250301205654/https://www.reddit.com/r/politics/hot/

[79] Reddit. 2025. *Archivedr/politics (by Hot, March 2025).* https://web.archive.org/web/20250301205654/https://www.reddit.com/r/politics/hot/

[80] Reddit. 2025. *Debunking Fukushima radiation fears: What tritium really means for ocean safety. (April 2025).* https://www.reddit.com/r/nuclear/comments/1k31ga5/debunking_fukushima_radiation_fears_what_tritium/

[81] Reddit. 2025. *r/transit (Arguments for Publicly Funding Rail and the Differences between Rail and Roads) (April 2025).* https://www.reddit.com/r/transit/comments/1jzix0j/arguments_for_publicly_funding_rail_and_the/

[82] Reddit Inc. 2025. *Introducing Reddit Answers.* https://redditinc.com/blog/introducing-reddit-answers/

[83] Elizabeth Reid. 2025. *Generative AI in Search: Let Google do the searching for you.* Google. https://blog.google/products/search/generative-ai-google-search-may-2024/

[84] Liam Reilly and Hadas Gold. 2025. *A 'rogue employee' was behind Grok's unprompted 'white genocide' mentions.* CNN. https://edition.cnn.com/2025/05/16/business/a-rogue-employee-was-behind-groks-unprompted-white-genocide-mentions

[85] Judy Robertson and Maurits Kaptein. 2016. *An Introduction to Modern Statistical Methods in HCI.* Springer International Publishing, Cham, 1–14. doi:10.1007/978-3-319-26633-6_1

[86] Jon Roozenbeek and Sander van der Linden. 2019. The fake news game: actively inoculating against the risk of misinformation. *Journal of Risk Research* 22, 5 (2019), 570–580. doi:10.1080/13669877.2018.1443491

[87] Todd Rose. 2022. *Collective Illusions: Conformity, Complicity, and the Science of why We Make Bad Decisions.* Hachette Books, New York, NY.

[88] Katja Rost, Lea Stahel, and Bruno S. Frey. 2016. Digital Social Norm Enforcement: Online Firestorms in Social Media. *PLOS ONE* 11, 6 (06 2016), 1–26. doi:10.1371/journal.pone.0155923

[89] Tahereh Saheb, Mouwafac Sidaoui, and Bill Schmarzo. 2024. Convergence of artificial intelligence with social media: A bibliometric & qualitative analysis. *Telematics and Informatics Reports* 14 (2024), 100146. doi:10.1016/j.teler.2024.100146

[90] Arthur D. Santana. 2014. Virtuous or Vitriolic. *Journalism Practice* 8, 1 (2014), 18–33. doi:10.1080/17512786.2013.813194

[91] Eric Hal Schwartz. 2024. *Reddit will use AI to summarize its insanely long threads and could transform the service.* Tech Radar. https://www.techradar.com/computing/artificial-intelligence/reddit-will-use-ai-to-summarize-its-insanely-long-threads-and-could-transform-the-service/

[92] Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. 2024. Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1033, 17 pages. doi:10.1145/3613904.3642459

[93] Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. 2024. CultureBank: An Online Community-Driven Knowledge Base Towards Culturally Aware Language Technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 4996–5025. doi:10.18653/v1/2024.findings-emnlp.288

[94] R. R. Silva, N. Chrobot, E. Newman, N. Schwarz, and S. Topolinski. 2017. Make It Short and Easy: Username Complexity Determines Trustworthiness Above and Beyond Objective Reputation. *Front Psychol* 8 (2017). doi:10.3389/fpsyg.2017.02200

[95] Jeff Smith, Alex Leavitt, and Grace Jackson. 2018. *Designing New Ways to Give Context to News Stories.* Facebook. https://about.fb.com/news/2018/04/inside-feed-article-context/

[96] Gregg Sparkman, Nathan Geiger, and Elke U. Weber. 2022. Americans experience a false social reality by underestimating popular climate policy support by nearly half. *Nature Communications* 13, 1 (2022), 4779. doi:10.1038/s41467-022-32412-y

[97] Dominik A. Stecula and Mark Pickup. 2021. Social Media, Cognitive Reflection, and Conspiracy Beliefs. *Frontiers in Political Science* 3 (2021). doi:10.3389/fpos.2021.647957

[98] Chris Stokel-Walker. 2025. *Reddit users were subjected to AI-powered experiment without consent.* New Scientist. https://www.newscientist.com/article/2478336-reddit-users-were-subjected-to-ai-powered-experiment-without-consent/

[99] J. Suler. 2004. The online disinhibition effect. *Cyberpsychol Behav* 7, 3 (2004). doi:10.1089/1094931041291295

[100] Henry Tari, M. Danial Khan, Justus Rutten, Darian Othman, Thales Bertaglia, Rishabh Kaushal, and Adriana Iamnitchi. 2024. Leveraging GPT for the Generation of Multi-Platform Social Media Datasets for Research. In *Proceedings of the 35th ACM Conference on Hypertext and Social Media* (Poznan, Poland)

(HT '24). Association for Computing Machinery, New York, NY, USA, 337–343. doi:10.1145/3648188.3675153

[101] Josh Taylor. 2025. *Australia's social media ban is attracting global praise – but we're no closer to knowing how it would work.* The Guardian. https://www.theguardian.com/technology/2025/apr/05/australia-social-media-ban-trial-global-response-implementation

[102] Josh Taylor. 2025. *Critics say Andrew Cuomo's transportation proposal for NYC sounds familiar.* NBCUniversal Media. https://www.nbcnewyork.com/new-york-city/andrew-cuomo-mayor-bus-plan/6177890/

[103] Josh Taylor. 2025. *'Zohran Mamdani represents the future New York': socialist riding high in bid to be mayor.* The Guardian. http://theguardian.com/us-news/2025/apr/19/zohran-mamdani-andrew-cuomo-new-york-city-mayor

[104] David R Thomas. 2006. A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation* 27, 2 (2006), 237–246.

[105] 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. doi:10.1111/j.2044-8309.1986.tb00732.x

[106] Rafal Urbaniak, Patrycja Tempska, Maria Dowgiałło, Michał Ptaszyński, Marcin Fortuna, Michał Marcińczuk, Jan Piesiewicz, Gniewosz Leliwa, Kamil Soliwoda, Ida Dziublewska, Nataliya Sulzhytskaya, Aleksandra Karnicka, Paweł Skrzek, Paula Karbowska, Maciej Brochocki, and Michał Wroczyński. 2022. Namespotting: Username toxicity and actual toxic behavior on Reddit. *Computers in Human Behavior* 136 (2022), 107371. doi:10.1016/j.chb.2022.107371

[107] Jenny S Wang, Samar Haider, Amir Tohidi, Anushkaa Gupta, Yuxuan Zhang, Chris Callison-Burch, David Rothschild, and Duncan J Watts. 2025. Media Bias Detector: Designing and Implementing a Tool for Real-Time Selection and Framing Bias Analysis in News Coverage. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 790, 27 pages. doi:10.1145/3706598.3713716

[108] Tzu-Yu Weng, Hanna Alzughbi, Isaac Rabago, Erin Arévalo Chaves, Erik Vagil, Nancy Fulda, Erin Ash, Mainack Mondal, Bart Knijnenburg, and Xinru Page. 2025. "Strangers in a new culture see only what they know": Evaluating Effectiveness of GPT-4 Omni for Detecting Cross-Cultural Communication Norm Violations. In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '25)*. Association for Computing Machinery, New York, NY, USA, 335–340. doi:10.1145/3699682.3728357

[109] Tim Weninger, Xihao Avi Zhu, and Jiawei Han. 2013. An exploration of discussion threads in social news sites: a case study of the Reddit community. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (Niagara, Ontario, Canada) *(ASONAM '13)*. Association for Computing Machinery, New York, NY, USA, 579–583. doi:10.1145/2492517.2492646

[110] Senuri Wijenayake, Danula Hettiachchi, Simo Hosio, Vassilis Kostakos, and Jorge Goncalves. 2021. Effect of Conformity on Perceived Trustworthiness of News in Social Media. *IEEE Internet Computing* 25, 1 (2021), 12–19. doi:10.1109/MIC.2020.3032410

[111] X Corp. 2025. *About Community Notes on X.* https://help.x.com/en/using-x/community-notes

[112] xAI [xAI]. 2025. *We want to update you on an incident that happened with our Grok response bot on X yesterday.* X Corp. https://x.com/xai/status/1923183620606619649

[113] Mari Yamaguchi. 2025. *'Nervous and rushed': Massive Fukushima plant cleanup work involves high radiation and stress.* https://apnews.com/article/japan-fukushima-plant-radiation-safety-4efe204a48f952137cac5a44b41f93ae

[114] Can Yang, Xinyuan Xu, Bernardo Pereira Nunes, and Sean Wolfgand Matsui Siqueira. 2023. Bubbles bursting: Investigating and measuring the personalisation of social media searches. *Telematics and Informatics* 82 (2023), 101999. doi:10.1016/j.tele.2023.101999

[115] 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. doi:10.1080/10584609.2020.1785067

[116] Koji Yatani. 2016. *Effect Sizes and Power Analysis in HCI.* Springer International Publishing, Cham, 87–110. doi:10.1007/978-3-319-26633-6_5

[117] YouTube. 2025. *Generative AI in Search: Let Google do the searching for you.* https://support.google.com/youtube/answer/14239105

[118] Yulin Yu, Julie Jiang, and Paramveer S. Dhillon. 2024. Characterizing the Structure of Online Conversations Across Reddit. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 374 (Nov. 2024), 23 pages. doi:10.1145/3686913

[119] Mawell Zeff. 2023. *Enter 'Grok,' Elon Musk's Anti-Woke Chatbot.* Gizmodo. https://gizmodo.com/enter-grok-elon-musks-anti-woke-chatbot-1850994584

[120] Yu Zhang, Jingwei Sun, Li Feng, Cen Yao, Mingming Fan, Liuxin Zhang, Qianying Wang, Xin Geng, and Yong Rui. 2024. See Widely, Think Wisely: Toward Designing a Generative Multi-agent System to Burst Filter Bubbles. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 484, 24 pages. doi:10.1145/3613904.3642545

A GPT4.5 Prompts Used

The following section includes the GPT-4.5 prompts used to generate the synthetic subreddit threads, as well as the AI narratives summary tool's responses. The AI percentages tool's agreement/disagreement percentages originate from the number of comments selected from the comment prompt (e.g., generating 10 pro-stance comments, compared to 20 anti-stance comments, and using this proportion to generate our percentages bars).

<TOPIC PREMISE> refers to the parent post's premise statement provided in subsection 3.1

A.1 AI Narratives Summary Tool Prompt

<TOPIC PREMISE>

<TOPIC COMMENTS>

Imagine you are a AI-bot that summarises the overall narratives (on the topics) and perspectives of the users arguing on the Reddit thread. Summarise the cases that supports the post (i.e., <AGREEMENT STANCE>) vs. opposes the stance of the post (i.e., <DISAGREEMENT STANCE>). Limit to around 50 words per stance and keep it polite and formal starting with "<Opponents, Supporters> believe that that...". Give percentages of how many comments support or oppose the topic's premise.

N.B. the final sentence acts as a sanity/manipulation check to ensure that our prompt could identify the stances in the comments provided to it to capture the correct stances to make its narrative summaries.

A.2 Comment Prompt

Create a synthetic reddit thread on the following debate topic: <TOPIC PREMISE>

In this experiment, there are two types of conversations: civil (fair and logical/evidence-driven argumentation and concessions across both sides, with evidence), and toxic (includes logical fallacies and ad hominem attacks, offensive statements, and profanity, with evidence). All posts should be realistic to how redditors speak. Keep each reddit comment to under 150 words each, leaning towards 50-100 words for each separate comment or reply.

Each reply must have two toxic comments where one must <SUPPORT, OPPOSE> and the other must <SUPPORT, OPPOSE> the commenter. Create <NUMBER OF COMMENTS TO GENERATE> main <TOXIC OR CIVIL TOPIC> comments. For each main comment, all comments/replies should be <TOXIC, CIVIL>. Give each comment a custom Reddit username.