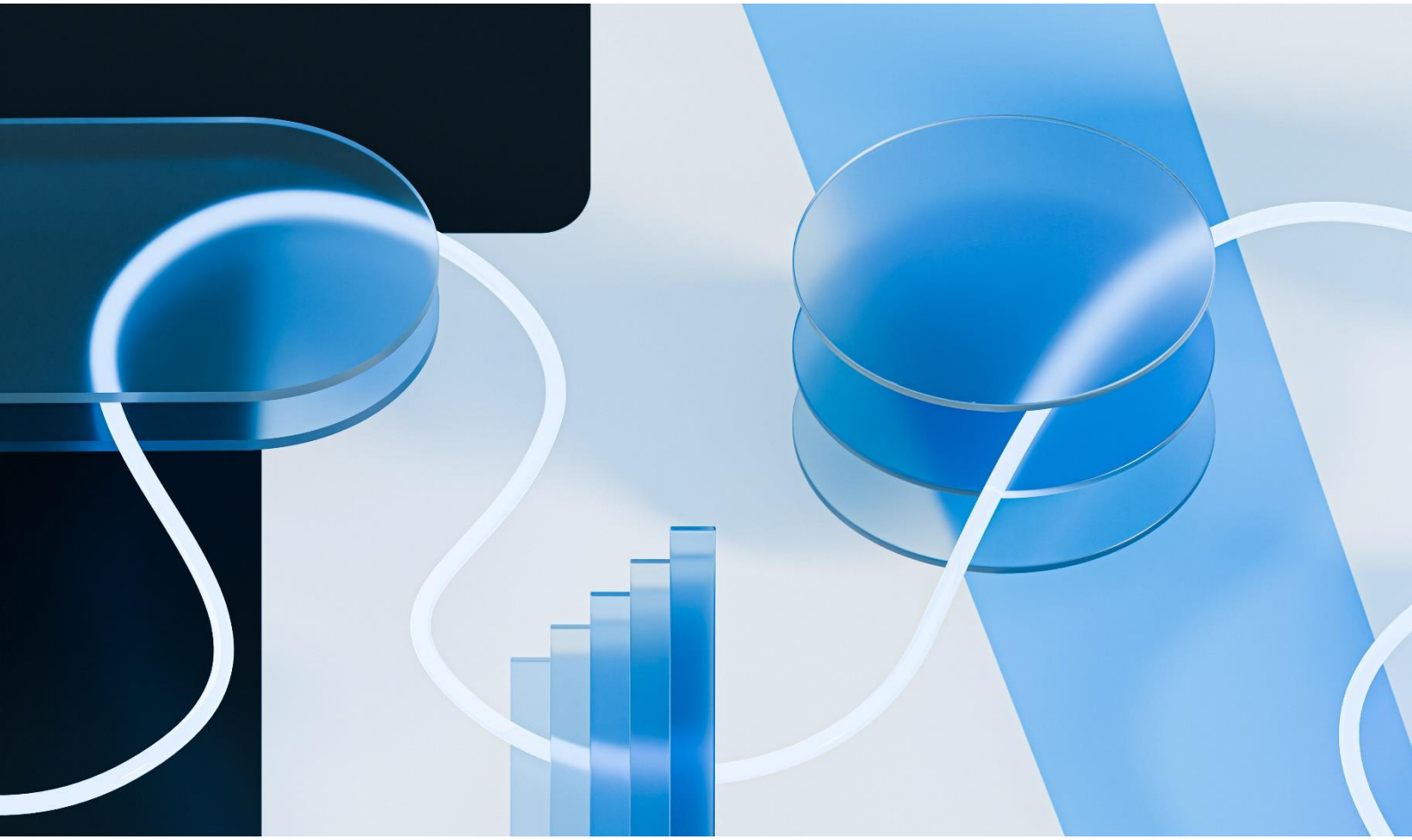
twon
twin of online social networks



TWin of Online Social Networks

Deliverable D5.3

Field Study Report

Main Authors: François t'Serstevens and Corinna Oschatz

Funded by
the European Union

About TWON

TWON (project number 101095095) is a research project, fully funded by the European Union, under the Horizon Europe framework (HORIZON-CL2-2022-DEMOCRACY-01, topic 07). TWON started on 1 April 2023 and will run until 31 March 2026. The project is coordinated by the Universiteit van Amsterdam (the Netherlands) and implemented together with partners from Universität Trier (Germany), Institut Jozef Stefan (Slovenia), FZI Forschungszentrum Informatik (Germany), Karlsruher Institut für Technologie (Germany), Robert Koch Institute (Germany), Univerzitet u Begogradu - Institut za Filozofiju I Drustvenu (Serbia) and Slovenska Tiskovna Agencija (Slovenia).

Funded by the European Union. Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

Project Name	Twin of Online Social Networks
Project Acronym	TWON
Project Number	101095095
Deliverable Number	D5.3
Deliverable Name	Field Study Report
Due Date	31.03.2025
Submission Date	27.03.2025
Type	R — Document, report
Dissemination Level	PU - Public
Work Package	WP 5
Lead beneficiary	1-UvA
Contributing beneficiaries and associated partners	Universität Trier (UT), FZI Forschungszentrum Informatik (Germany), Karlsruher Institut für Technologie (KIT), Robert Koch Institut (RKI), Univerzitet u Begogradu - Institut za Filozofiju I Drustvenu (UoB), Slovenska Tiskovna Agencija (STA), Dialogue Perspectives e.V (DIA)

Executive Summary

This report presents key findings from the TWON project's experimental research into online social network dynamics, contributing empirical evidence across three policy-relevant domains.

A central methodological contribution of this work is the use of a fully functioning simulated social media platform, enabling testing of interventions under ecologically valid conditions that closely mirror real-world information environments. This approach consistently revealed that effect sizes established in traditional laboratory or vignette-based studies do not straightforwardly translate to realistic settings, where interventions must compete for attention within a continuous stream of diverse content. The interventions examined were of direct interest to key public health and media institutions, including the Robert Koch Institute (RKI) and the Slovenian Press Agency (STA).

Beyond replicating existing platform dynamics, TWON also enables testing of features that do not yet exist on real social media. The third study demonstrates this potential, showing that opt-in AI-mediated paraphrasing can reduce incivility with spillover effects across entire conversation threads — an intervention that could not be studied without a purpose-built experimental platform.

Finally, we replicate the settings of our case studies in simulation, allowing us to examine how the tested interventions affect online social networks systematically and at scale. This reveals social media dynamics that controlled experiments with individual participants cannot readily capture, and points toward a fully automated research infrastructure for future policy-relevant investigation.

List of Tables

1.1 Participant Demographics 18
1.2 Logistic Regression Results 21

2.1 Sample characteristics by condition 30

3.1 Sample Demographics by Treatment Response 48
3.2 Comparisons of Toxicity Scores Between Original and Posted Comments 51
3.3 Hierarchical Regression Results of Conversation-Level Toxicity 54

4.1 Debate quality metrics by experimental condition 60

List of Figures

1.1 Example of Feed 19

2.1 Research Framework 27
2.2 Generic post within Studies 1 and 2 31
2.3 Distributions of Trust in RKI and Post accuracy across experiments and conditions . . . 32
2.4 Uncertainty Framing Stimuli 35
2.5 Structural Equation Model of the Uncertainty Framing Study 35
2.6 Prebunking Control 38
2.7 Prebunking Treatment 38
2.8 Structural Equation Model of the Prebunking Study 38

3.1 Example of a post in the TWON Feed 46
3.2 Differences in toxicity between original and paraphrased comments 49
3.3 Posterior distributions and 95% credible intervals for key regression coefficients. 53

Contents

List of Tables **5**

List of Figures **6**

List of Abbreviations **9**

I Case Studies **12**

1 Recommender Systems and Misinformation Belief **15**

1.1 Introduction . . . 16

1.2 Theoretical Framing . . . 16

1.3 Experimental Procedure . . . 17

1.4 Results . . . 20

1.5 Discussion . . . 22

2 Effectiveness of Prebunking and Uncertainty Frames in Vaccine Messaging **24**

2.1 Introduction . . . 25

2.2 Communicating Scientific Information . . . 26

2.2.1 Uncertainty Communication . . . 26

2.2.2 Prebunking . . . 28

2.3 Experimental Design . . . 29

2.4 Study 1: Uncertainty Framing . . . 33

2.5 Study 2: Prebunking . . . 34

2.6 Discussion . . . 37

3 Nudging civility on OSN Using LLMs **41**

3.1 Introduction . . . 42
3.1.1 Toxicity in OSN . . . 43
3.1.2 Common Approaches . . . 43
3.1.3 Generative AI as an intervention . . . 44
3.2 Methods . . . 45
3.2.1 Experimental Design and Platform . . . 45
3.2.2 Procedure . . . 45
3.3 Analysis and Results . . . 47
3.3.1 LLMs and Toxicity reduction . . . 48
3.4 Discussion & Conclusion . . . 55

II Simulation studies **56**

4 Simulation Debate Quality **58**
4.1 Introduction . . . 58
4.2 Simulation Detail . . . 59
4.2.1 Results and Discussion . . . 60

5 Towards simulating Serbian debates **62**
5.1 Introduction . . . 62
5.2 A first LLM-based simulation . . . 62
5.3 Calibrating LLM-agents on Serbian survey respondents . . . 63

Bibliography **67**

List of Abbreviations

BERT	Bidirectional Encoder Representations from Transformers
CFA	Confirmatory Factor Analysis
GLLM	Generative Large Language Model
LLM	Large Language Model
mRNA	Messenger ribonucleic acid
OSN	Online Social Network
RKI	Robert-Koch-Institute
SEM	Structural Equation Modelling
STA	Slovenian Press Agency
TWON	Twin of Online Social Networks
UT	University of Trier
UvA	University of Amsterdam

Introduction

Developing a digital twin of an online social network, a TWON, can be seen as an iterative process: The development of the technical infrastructure and the fitting and calibration of models and agents to at least some extent relies on empirical data; versions of the TWON can be tested against empirical data, and ultimately, the TWON can be used to predict user behaviour on OSN.

Hence, within the TWON project, we from the outset planned to conduct a variety of case studies, centering around the themes of health communication (in particular, around pandemics and vaccination) and communication in times of war (in particular, the war of Russia against Ukraine). These studies were planned to serve multiple goals:

Address substantive questions by societal partners. Developing a TWON is not purely an academic exercise. We in particular wanted to answer substantive questions that non-academic, societal partners have. In particular, the Robert-Koch-Institute (RKI), a German federal government agency responsible for disease control, was asking in how far a TWON could be used to test two communication strategies that are frequently used in a health-communication context; the so-called truth-sandwich and prebunking (Chapter 2). STA, a Slovenian press agency, was wondering in how far different platform mechanics – in particular, recommender systems that either featured re-inforcing or opposing content – could explain engagement with disinformation the context of war (Chapter 1). Beyond the demands of our partners, we also report on an overarching case in which we aimed at making the link between platform mechanics and debate quality more explicit, by exploring new features that may reduce toxicity in online discussions (Chapter 3).

Establish new methods. TWONs goal is to ultimately produce fully-automated simulations (see, for instance, Part II of this report). Nonetheless, it is important to understand the nature of human behaviour before we replicate it. We therefore specifically aimed at exploring human behaviour

on a platform that is as realistic as possible, either only amongst humans, or between humans and bots. In this sense, the case studies also served as a way to iteratively improve ways of studying platform mechanics.

Generate data for machine learning. During the project duration, the state-of-the-art changed considerably, implying a shift towards LLM-based approaches for creating agents – which changes the needs for training data. We have shown within the TWON-project that careful evaluation on real-world data is still necessary (Münker et al., 2025). For this purpose, we could rely on data we collected previously (Schwager et al., 2026; Münker et al., 2026), and which we described earlier in deliverables D3.2 and D4.2. We therefore will not reiterate these in this deliverable. The collection of additional training data via hybrid experiments with humans using the TWON-user interface (see also deliverable D4.3) was therefore deprioritised.

This report, therefore, is divided into two parts. First, we discuss specific case studies we conducted involving human participants. In the second part, we report how these can be connected to fully-automated simulations.

Part I

Case Studies

Case Studies Overview

Part I describes the case studies conducted to test and further develop the TWON platform and gather user data on problematic online behavior.

Case study 1 investigated the role of recommender systems in driving user engagement with disinformation in the context of high-stakes geopolitical conflicts. This study was conducted together with our partner, the Slovenian Press Agency (STA), using the Russia-Ukraine war as current example. Participants from Serbia interacted with news content related to the Russia–Ukraine war that systematically varied factual accuracy and political perspective (pro-Russian, pro-Ukrainian, neutral). Content feeds were generated under three algorithmic conditions: reinforcing, opposing, and no-curation control. This study also constituted the first full deployment of the TWON platform, through which the infrastructure was substantially refined and stabilised.

Case study 2 was conducted in close collaboration with the Robert-Koch-Institut (RKI), which is Germany's leading agency for disease surveillance and prevention. One of the main problems the RKI currently encounters is a decreasing trust in scientific institutions and evidence in vaccination among citizens due to a large influx of false information surrounding vaccines. This case study aims to tackle these challenges by exploring the effectiveness of two communication strategies in a combined experiment. We use the TWON platform to examine the effectiveness of prebunking and uncertainty-framed messaging on public attitudes toward vaccines and trust in the scientific community. Participants recruited through Prolific, an online recruiting platform, navigated a feed of diverse posts, with one post experimentally manipulated to include either (i) prebunking content, i.e. a strategy that warns the audience of potentially false and misleading information about a topic, or (ii) an uncertainty frame regarding vaccine effectiveness. Results provide no evidence that either intervention increased the trust in science or the post itself. Whereas vignette studies have suggested these methods can be effective. We argue that this is due to the ecological nature of our experiment.

In study 3, we use a novel application of LLMs to counter uncivil and toxic interactions of social media platforms before they occur. We employ recent advances in generative AI to study whether AI-mediated paraphrasing interventions can reduce uncivil expression while preserving substantive disagreement in online political discussions, which is a normal and required feature of democratic debates. Using TWON, we analyse how AI-generated paraphrases influence the tone of conversations and assess their effects not only on the direct participants of a debate but also on external observers who encounter these exchanges. The findings provide insights into the potential of AI-assisted communication tools to foster healthier online discourse at scale.

Chapter 1

Recommender Systems and Misinformation Belief

Abstract

Social media recommender systems shape the information users encounter and influence how they engage with it, potentially affecting knowledge formation, attitudes, and behavior. At the same time, concerns about the spread of disinformation on social media platforms have grown substantially in the past decade. However, the role of recommender systems in driving user engagement with disinformation remains insufficiently understood, particularly in the context of high-stakes geopolitical conflicts. This study addresses this gap using the TWON platform, an experimental social media environment that allows manipulation of algorithmic curation under ecologically valid conditions. Participants from Serbia interacted with news content related to the Russia–Ukraine war that systematically varied factual accuracy and political perspective (pro-Russian, pro-Ukrainian, neutral). Content feeds were generated under three algorithmic conditions: reinforcing, opposing, and no-curation control. Contrary to common assumptions, we find no consistent effect of recommender system type on engagement with disinformation. Instead, participants were generally less likely to engage with highly distorted content regardless of algorithmic condition. These findings suggest that algorithmic curation alone may not sufficiently explain the spread of disinformation and highlight the importance of user-level responses to content quality.

1.1 Introduction

Social media recommender systems play a decisive role in determining what content users encounter and how they engage with it. The engagement with this information, in turn, substantially determines what knowledge users gain, what opinions they form and what attitudes and behaviors they develop. At the same time, we observe a substantial increase in the spread of false and inaccurate information referred to as disinformation, misinformation or generally fake news on the social media platforms in the past decade (Aïmeur et al., 2023; Lecheler and Egelhofer, 2022; Broda and Strömbäck, 2024). Yet, the impact of recommender systems on user engagement with fake news remains poorly understood, particularly in the context of high-stakes geopolitical conflicts where competing narratives circulate widely. We address this gap using the TWON platform that enables us to test algorithmic curation of recommender systems in a more natural setting. This ecologically valid setting is central to our approach, rather than relying on observational data. We place participants inside a functioning platform that replicates popular OSNs, allowing us to manipulate the recommender system design. The content used in the study was developed in collaboration with the Slovenian Press Agency (STA), which provided us with news posts based on real reporting and systematically varied the degree of disinformation, from factually accurate to heavily distorted.

In this study, we specifically focus on one platform mechanic: the effect of recommender systems on engagement with disinformation in the context of the war between Russia and Ukraine. Participants from Serbia interact with content varying in political perspective (pro-Russian, pro-Ukrainian, neutral) and factual accuracy, under three algorithmic conditions: a reinforcing algorithm that serves content aligned with a user's prior engagement, an opposing algorithm that introduces counter-attitudinal content, and a no-curation control. *We ask to what extent the recommender system type determines user engagement with disinformation*. Our findings offer empirical grounding for ongoing debates on platform design and the regulation of online disinformation.

1.2 Theoretical Framing

Recommender systems are designed to maximize user engagement by surfacing content predicted to elicit interaction (Sunstein, 2018; Pariser, 2011). In doing so, they act as powerful gatekeepers of the information environment, shaping not only what users see but how much they interact with it (Ribeiro et al., 2020). Research consistently shows that algorithmically curated feeds increase time-on-platform and interaction rates relative to chronological or other alternatives (Bail et al., 2018; Covington

et al., 2016). Crucially, engagement-maximizing algorithms tend to favour content that is emotionally arousing-a common characteristic of disinformation as it is frequently crafted to provoke strong affective responses (Vosoughi et al., 2018).

Reinforcing algorithms, by continuously providing content similar to content a user has shown engagement with previously, likely contribute to the formation of echo chambers and filter bubbles (Pariser, 2011; Flaxman et al., 2016). Within these environments, disinformation that "resonates" with existing beliefs faces little friction: it is engaged with, shared, and further amplified, creating a feedback loop in which ideologically congruent but factually questionable content accrues disproportionate visibility (Guess et al., 2019; Pennycook and Rand, 2021). Yet, the relationship between algorithmic exposure and belief alteration is not straightforward. When users encounter counter-attitudinal content, the response is often defensive rather than deliberative: disengagement, negative reaction, or reinforcement of prior beliefs (Bail et al., 2018; Nyhan and Reifler, 2010; Guess and Lyons, 2020). Reinforcing and opposing algorithms thus operate differently not only in what they put forward, but also in how users engage with it.

These dynamics are well-theorized but difficult to study on live platforms, where algorithmic conditions cannot be isolated and user behaviour is confounded by platform-wide effects. The relative impact of reinforcing versus opposing recommendation on engagement with disinformation therefore remains an open empirical question, and one with direct implications for how platforms are designed and regulated. Our study addresses it under controlled experimental conditions.

1.3 Experimental Procedure

A total of 284 Serbian participants were recruited to take part in our study. From the original sample, $N = 189$ completed all rounds and are included in the analysis. Participants were randomly assigned to one of three algorithmic conditions, resulting in three roughly equal groups (73 for *control*, 62 for *opposing*, and 54 for *reinforcing*). Table 1.1 describes the sample at hand. The study followed a single-session between-subjects design. Participants first completed a pre-survey measuring baseline political attitudes, demographic information, and prior knowledge about the Ukraine conflict. They were then directed to the TWON platform for the treatment phase, after which they completed a post-survey re-assessing content evaluations and attitudinal outcomes.

The experiment was conducted on the TWON platform where participants browsed a feed of posts and could interact with content through likes and reactions. An example of the platform feed as seen by par-

Table 1.1: Participant Demographics

Variable	**Category**	***n* (%)**
Age	*Mean* (SD)	45.77 (12.33)
Gender	Man	126 (66.7%)
	Woman	63 (33.3%)
Education	Basic/secondary	77 (40.7%)
	Secondary	97 (51.3%)
	Other	15 (7.9%)
Employment	Employed	122 (64.6%)

Note. $N = 189$.

ticipants is shown in Figure 1.1. The feed consisted of two types of posts. All posts, and particularly, the primary stimuli, were news posts developed in collaboration with the Slovenian Press Agency, based on real reporting on the war between Russia and Ukraine. Each article was rewritten into variants along two dimensions: political perspective (pro-Russian, pro-Ukrainian, and neutral) and degree of disinformation (ranging from factually accurate to heavily distorted). Alongside the stimuli, participants also encountered a set of filler posts unrelated to the conflict — covering everyday topics such as sports and non-Ukraine-based political news as they would be encountered in participants' regular social media feeds creating a naturalistic scrolling experience. Half of the filler and stimuli posts featured a full article that was accessible in-platform[1].

Algorithmic personalisation was based on participants' in-platform behaviour: likes and reactions to posts were used to infer orientation and to update the feed. In the *reinforcing* condition, the algorithm prioritized content aligned with each participant's inferred positioning. In the *opposing* condition, the platform pushed counter-attitudinal content forward instead. In the *control* condition, no personalisation was applied and posts were served without algorithmic curation.

The feed was structured in three rounds with all participants starting from a randomly ordered set of posts in round 1. Each Ukraine-related post was assigned a `DisinfoScore` reflecting its degree of factual distortion (0 for Objective, 1 for light, and 2 for heavy disinformation). Participants interacted with

[1] For stimuli posts, the article was rewritten according to the intended disinformation level.

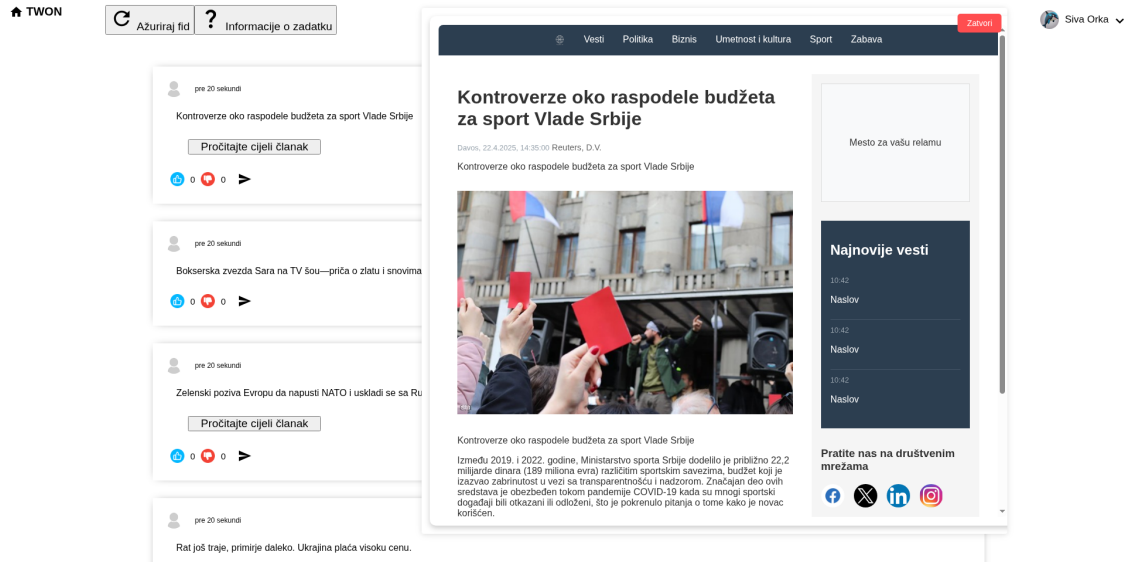


Figure 1.1: Example of Feed

posts through views, likes, dislikes, and comments. Between rounds, the platform computed an engagement score for each post with which a participant had interacted. These scores were then aggregated across all posts to derive user-level alignment scores. The formal procedure is shown in Equations 1.1 and 1.2.

$$\text{InteractionScore} = \text{views} + \text{likes} - \text{dislikes} + 0.5 \times \text{comments} \tag{1.1}$$

$$\text{OrderScore}_i = \sum_{p \in P_i} \text{InteractionScore}_p \times \text{DisinfoScore}_p \tag{1.2}$$

Where p indexes posts and P_i denotes the set of posts seen in round i.

The `OrderScore` determined the composition of the feed in the subsequent round: in the *reinforcing* condition, posts consistent with the user's inferred position were prioritised; in the *opposing* condition, counter-attitudinal posts were up-ranked. In the *control* condition, the feed composition remained randomized throughout. Importantly, the feed was not updated within a round but recalculated at fixed transition points, such that likes in round 1 determined the feed in round 2, and likes in round 2 determined the feed in round 3. The three-round structure reflects two practical constraints: a minimum of two rounds is required for the recommendation algorithm to adapt based on prior engagement, and additional rounds incur substantial preparation costs per article set.

1.4 Results

We estimated three logistic regression models predicting whether a participant liked, disliked, or commented on a post, with the disinformation level (1 = factual through 3 = heavy disinformation) and algorithmic condition (reference: control) as predictors, including their interaction. Results are presented in Table 1.2. Overall, the results offer little evidence that the recommender system type systematically shapes engagement with disinformation. Neither the opposing nor the reinforcing condition produced a consistent pattern of effects across engagement types, and the majority of interaction terms between algorithmic condition and disinformation level are non-significant. Where significant effects did emerge, they are scattered across outcomes and disinformation levels, and do not show a clear pattern of algorithmic amplification or suppression. This holds for both conditions: the opposing algorithm does not consistently dampen engagement with disinformation, nor does the reinforcing algorithm consistently inflate it.

The main effects of the disinformation level suggest that participants were sensitive to content quality independently of the algorithmic condition. Heavy disinformation attracted fewer likes ($b = -0.711$, $p < .05$) and more dislikes ($b = 0.881$, $p < .05$) than factual content. Likewise, moderate disinformation attracted more dislikes ($b = 1.258$, $p < .01$) than factual content. This baseline tendency to react negatively to more distorted content is notable on its own as it indicates that participants were not indifferent to the disinformation level when engaging with posts, even in the absence of algorithmic curation.

The main effects of algorithmic condition are also worth noting, though they should be interpreted with care. Relative to the control, the reinforcing condition was associated with lower liking ($b = -0.312$, $p < .05$), more dislikes ($b = 0.777$, $p < .01$), and less commenting ($b = -0.781$, $p < .05$). However, since these are effects independent of disinformation level, and given that the corresponding interaction terms are largely non-significant, they do not indicate that the reinforcing algorithm specifically altered how participants engaged with disinformation content as opposed to posts more generally. The one exception is a weakly significant increase in commenting on heavy disinformation in the reinforcing condition ($b = 2.225$, $p < .10$), though this should be interpreted cautiously given the very low base rate of comments and the highly unstable estimates for that outcome, as reflected in the extreme standard error for the moderate disinformation interaction term.

The opposing condition interactions lack coherence. Slight disinformation attracted more likes ($b = 1.042, p < .01$) and heavy disinformation marginally more likes ($b = 0.720, p < .10$) relative to the con-

trol, while moderate disinformation attracted fewer dislikes ($b = -1.173$, $p < .05$). These effects point in different directions across disinformation levels and do not support a single interpretation. Taken together with the reinforcing condition results, the overall picture is one of noise rather than signal: recommender system type did not meaningfully determine how participants engaged with disinformation in this experimental setting.

Table 1.2: Logistic Regression Results

	Dependent variable:		
	Liked	Disliked	Commented
	(1)	(2)	(3)
Disinfo: slight (1)	−0.465	0.515	0.039
	(0.289)	(0.405)	(0.623)
Disinfo: moderate (2)	−0.113	1.258***	0.039
	(0.268)	(0.328)	(0.623)
Disinfo: heavy (3)	−0.711**	0.881**	−1.087
	(0.309)	(0.362)	(1.027)
Opposing	−0.175	0.292	−0.471
	(0.125)	(0.214)	(0.337)
Reinforcing	−0.312**	0.777***	−0.781**
	(0.133)	(0.203)	(0.391)
Disinfo slight × Opposing	1.042***	−0.430	0.219
	(0.398)	(0.609)	(0.989)
Disinfo moderate × Opposing	0.122	−1.173**	0.642
	(0.399)	(0.561)	(0.901)
Disinfo heavy × Opposing	0.720*	−0.996	0.636
	(0.427)	(0.609)	(1.464)
Disinfo slight × Reinforcing	0.355	0.268	−0.039
	(0.442)	(0.527)	(1.233)
Disinfo moderate × Reinforcing	0.197	−0.788	−13.635
	(0.418)	(0.489)	(538.368)
Disinfo heavy × Reinforcing	0.141	−0.003	2.225*
	(0.489)	(0.490)	(1.233)
Constant	−0.728***	−2.610***	−3.189***
	(0.083)	(0.154)	(0.200)

Continued on next page

	Liked	Disliked	Commented
	(1)	(2)	(3)
Observations	2,268	2,268	2,268
Log Likelihood	−1,351.961	−767.455	−292.524
Akaike Inf. Crit.	2,727.922	1,558.910	609.047

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$
Reference categories: factual content (disinfo label 0), control condition.

1.5 Discussion

The results of this study add to a growing body of work questioning the assumption that recommender systems are straightforward drivers of disinformation engagement (Guess et al., 2019, 2018). Neither reinforcing nor opposing algorithms produced a consistent effect on how participants engaged with disinformation content, suggesting that the type of algorithmic curation alone may be insufficient to explain the spread of false or distorted information online. This is a meaningful finding in its own right. Much of the policy debate around platform regulation has centred on the role of recommendation in amplifying disinformation (Sunstein, 2018; Pariser, 2011; Nguyen, 2020), yet our findings suggest the relationship is considerably more complex. Participants responded to content quality — disliking more distorted posts regardless of condition — which points to a degree of critical engagement that purely algorithmic accounts of disinformation spread tend to underestimate (Pennycook and Rand, 2021; Guess et al., 2018, 2019). At the same time, the absence of strong interaction effects does not mean recommender systems are irrelevant; it may instead reflect the limitations of the current study.

The platform represents a methodological advancement in the study of social media behaviour. By placing participants inside a functioning platform, we are able to manipulate algorithmic conditions in a setting that is ecologically grounded (Salganik et al., 2006; Bronfenbrenner, 1977) something that is not possible with observational data from real platforms, where conditions are confounded and algorithms (Ribeiro et al., 2020). From a usability standpoint, the platform performed well: participants navigated the interface and engaged with posts without apparent difficulty, lending confidence that the experimental environment was sufficiently naturalistic to elicit genuine engagement behaviour. This is a non-trivial achievement for a purpose-built research tool and an important proof of concept for the broader TWON research program.

That said, the current platform design has real limitations that bear on the interpretation of the results. Most notably, the recommender algorithm implemented in TWON is a simplified approximation of sophisticated algorithms by major platforms (Covington et al., 2016; Chen et al., 2019). Real-world algorithms draw on extensive behavioural histories, social graph data, and optimization objectives that cannot be replicated in a short experimental session with a fresh user base (Ribeiro et al., 2020). The three-round feed update structure, while a reasonable operationalization, provides only a coarse simulation of the continuous, dynamic curation that characterises platforms like Twitter/X or Facebook. Relatedly, the sample size, while adequate for a pilot study, is modest relative to the scale at which real platform dynamics operate (Bail et al., 2018). Firstly, this limits the statistical power available to detect smaller interaction effects. Secondly, and more impactful for follow-up studies, the TWON platform essentially still presents end-users with a survey. Users do not have the same attachments that they would have for their preferred social media platforms. Therefore, the results of this study or similar studies are bound to be proxies of real-life behaviour.

Chapter 2

Effectiveness of Prebunking and Uncertainty Frames in Vaccine Messaging

Abstract

In this case study, we use the TWON platform to examine the effectiveness of prebunking and uncertainty-framed messaging on public attitudes toward vaccines and trust in the scientific community, together with our project partner RKI. Participants recruited through Prolific Academic navigated a feed of diverse posts spanning multiple topics, with one post experimentally manipulated to include either prebunking content, i.e. a strategy that warns the audience of potentially false and misleading information about a topic, or an uncertainty frame regarding vaccine effectiveness. Following interaction with the feed, respondents completed a survey assessing perceptions of the posts as well as general trust in science and institutions. The results of our case study provide no evidence that either intervention increased trust in science or the post itself. Whereas vignette studies have suggested these methods can be effective. We argue that this is due to the ecological nature of our experiment. Beyond the substantive findings, this study highlights the methodological value of using interactive, ecologically valid platforms to test health communication strategies, offering insights for both theory and applied messaging design.

2.1 Introduction

Vaccine hesitancy remains a persistent public health challenge, often fueled by misinformation circulating on social media (Roozenbeek et al., 2020). Understanding how to effectively communicate accurate health information is critical not only for promoting vaccination uptake but also for maintaining public trust in scientific institutions. Traditional studies on health messaging have typically relied on surveys or vignette-based experiments, which expose participants to health content in controlled isolation and measure their responses (Kreps and Kriner, 2020; Nan and Madden, 2012; van der Linden and Leiserowitz, 2019). Yet, the conditions under which people actually encounter health information online differ substantially from those created in laboratory settings. Real-world social media users are exposed to health messages embedded within a continuous stream of competing content, ranging from entertainment and news to interpersonal communication, and this broader informational context is likely to shape how messages are processed and evaluated (Guess and Lyons, 2020). This section of the report seeks to bridge that gap by testing the effects of two theoretically grounded messaging strategies, prebunking, i.e. a strategy that warns audience of potentially false and misleading information about a topic and uncertainty framing, within a controlled yet realistic digital environment that closely mirrors contemporary social media platforms.

These two strategies were selected because they represent distinct theoretical approaches to the same underlying problem of how to reduce the influence of health misinformation. Prebunking draws on inoculation theory, exposing individuals to a weakened form of misinformation alongside explicit refutations in order to build cognitive resistance before organic exposure occurs (Lewandowsky and van der Linden, 2021; van der Linden et al., 2017). Within this framework, we examine a so-called “truth sandwich”, a specific message structure that presents accurate information both before and after a brief acknowledgement of a false claim, minimizing the persuasive footprint of misinformation while reinforcing the corrective message (Amazeen et al., 2018). The second strategy, uncertainty framing, takes a complementary approach: instead of directly countering false claims, it communicates the limits or evolving nature of scientific knowledge in an effort to foster transparency and epistemic trust (van der Linden and Leiserowitz, 2019; Budescu et al., 2021). Despite the theoretical appeal of both approaches, evidence of their effectiveness remains mixed, particularly in ecologically valid settings where interventions must compete with the full complexity of a real information feed (Tulin et al., 2025; Roozenbeek and Van der Linden, 2019; Kreps and Kriner, 2020).

The present study was developed in direct partnership with the Robert Koch Institute (RKI), Germany’s

primary public health authority, which co-designed the experimental stimuli to ensure that all health communication content met institutional standards for accuracy and responsible messaging. Participants navigated on the TWON platform, encountering a feed of diverse posts spanning politics, entertainment, and health, with one post being experimentally manipulated to deliver either prebunking content in a truth sandwich variant, or an uncertainty frame regarding vaccine effectiveness. By embedding the experimental intervention within a realistic multi-topic feed rather than presenting it in isolation, we evaluate messaging strategies under conditions that more closely approximate real-world information exposure. We therefore expect that effect sizes will be more conservative than those reported in vignette-based studies, and we examine whether either strategy is sufficient to shift vaccine attitudes and intentions when competing for attention within a broader informational environment.

2.2 Communicating Scientific Information

Health communication research has long grappled with how best to convey scientific findings to a public that is simultaneously overwhelmed with competing information and increasingly sceptical of institutional authority (Lewandowsky and van der Linden, 2021). Central to this challenge are the concepts of perceived source and message credibility, i.e. the degree to which an audience judges a message and its author to be reliable, accurate, and trustworthy (van der Linden and Leiserowitz, 2019). A substantial body of research suggests that credibility perceptions function as a gateway through which health messages must pass before they can influence deeper attitudinal outcomes such as institutional trust or behavioural intentions (Kreps and Kriner, 2020). When a message is perceived as credible, audiences are more likely to accept its content and extend trust to its source. On the contrary, when credibility is undermined, even accurate and well-intentioned messages may fail to produce their intended effects (Kreps and Kriner, 2020; Roozenbeek et al., 2020). Figure 2.1 summarises the hypothesis developed in this section.

2.2.1 Uncertainty Communication

Communicating uncertainty involves explicitly acknowledging the limits, variability, or evolving nature of scientific knowledge. In the context of vaccines and other public health interventions, uncertainty can be conveyed through ranges of effectiveness, confidence intervals, or qualifying statements about ongoing research. Theoretical work suggests that transparency about uncertainty can increase credibility and foster trust, particularly among audiences who value honesty and scientific integrity (van der Linden and Leiserowitz, 2019). By signalling that scientists, or the authorities, are cautious

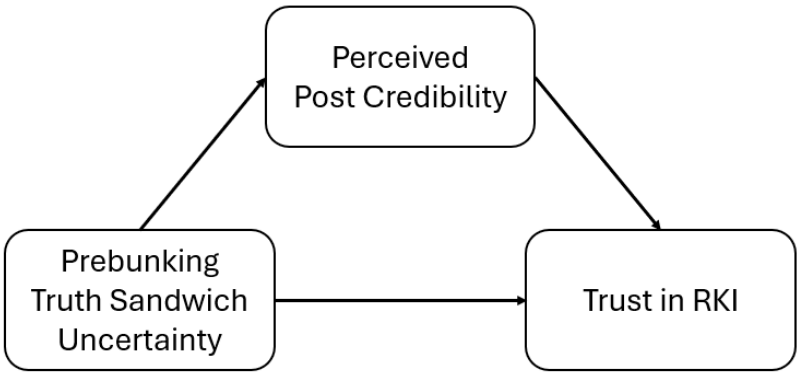


Figure 2.1: Research Framework

and evidence-driven rather than overconfident, uncertainty framing may strengthen long-term confidence in the scientific community.

Despite its theoretical grounding, empirical studies have found mixed results for uncertainty communication (Gustafson and Rice, 2020, 2019). On the positive side, acknowledging uncertainty can mitigate perceptions of deception or overstatement when outcomes do not match initial predictions (Budescu et al., 2021). Transparency in messaging has also been linked to increased engagement with scientific information and greater willingness to seek additional knowledge, suggesting that audiences appreciate nuance and careful reasoning. When combined with clear actionable guidance, uncertainty framing can maintain trust while still supporting health-promoting behaviour, suggesting that audiences are capable of holding nuance alongside motivation to act (Kreps and Kriner, 2020; Gustafson and Rice, 2020).

However, communicating uncertainty carries potential drawbacks. Explicitly highlighting uncertainty can inadvertently reduce perceived effectiveness, reliability, or confidence in recommended interventions, particularly among audiences with lower scientific literacy or stronger preexisting skepticism (Kreps and Kriner, 2020). Some studies suggest that when people interpret uncertainty as incompetence or indecision, it may decrease compliance or lead to increased susceptibility to misinformation (Roozenbeek et al., 2020). This tension between transparency as a credibility signal and uncertainty as a credibility threat is particularly likely to manifest in ecologically valid settings, where uncertainty cues must compete with other content and where audiences bring varied prior orientations to the same message (Guess and Lyons, 2020).

To the extent that uncertainty framing succeeds in signaling transparency and careful reasoning, it is expected to enhance perceived post credibility, which should in turn increase trust in the RKI.

2.2.2 Prebunking

The technique of prebunking, a form of inoculation, makes individuals aware of false information before they are organically exposed to it, thus staving off receptivity to the misinformation in advance. Rooted in inoculation theory (McGuire, 1961), prebunking consists of exposing people to misleading claims alongside refutations or counterarguments, building cognitive resistance to subsequent misinformation. This approach has been found effective across a variety of domains, including politics, climate change, and public health (Lewandowsky and van der Linden, 2021; van der Linden et al., 2017). The protective mechanism operates through two pathways: (i) motivational, by alerting individuals to the threat of manipulation, and (ii) cognitive, by pre-equipping them with counterarguments (van der Linden et al., 2017). In doing so, prebunking actively reinforces the perceived reliability of corrective sources, positioning the communicator as transparent and trustworthy because they acknowledge and address misinformation directly (Lewandowsky and van der Linden, 2021). Within the healthcare context, prebunking has shown promise in countering vaccine misinformation, increasing individuals' ability to recognise and reject false claims while sustaining trust in scientific guidance (Roozenbeek et al., 2020). Importantly for our purposes, the mechanism through which these effects emerge is credibility-dependent. Prebunking works, in part, because the source is seen as forthcoming, and that forthrightness translates into perceived trustworthiness (Basol et al., 2021).

Despite its potential, the effectiveness of prebunking depends on several factors, including message framing, timing, and audience characteristics. Research suggests that prebunking is most effective when interventions are brief, targeted, and easily comprehensible, allowing individuals to internalise the corrective information without cognitive overload (Basol et al., 2021). Moreover, the social and informational context in which prebunking occurs can moderate its impact. That is, when misinformation is embedded within highly engaging social feeds, the protective effect may be weakened if competing content distracts or overwhelms the audience (Guess and Lyons, 2020; Pennycook et al., 2020). These findings highlight the importance of testing prebunking interventions in ecologically valid environments that approximate the complex information ecosystems of modern social media.

Prebunking also carries the potential for unintended consequences. By exposing individuals to misinformation in advance, there is a risk that some recipients may focus on the misleading content rather than the corrective message, potentially increasing interest in or spreading the misinformation (Pennycook and Rand, 2020). Moreover, resistance to prebunking may be heterogeneous, with it being particularly pronounced among audiences with strong preexisting beliefs or low trust in the source delivering the intervention, or Science in general (Roozenbeek and Van der Linden, 2019). These considerations

underscore the need for careful message design and empirical evaluation of prebunking strategies, particularly in realistic settings where information is consumed alongside diverse and competing content.

A specific variant of prebunking that has gained traction in applied health communication is the truth sandwich (Tulin et al., 2025; Amazeen et al., 2018). Rather than presenting a correction after misinformation has been introduced, the truth sandwich structure starts with accurate information, briefly acknowledges the false claim, and then reaffirms the truth. The rationale is cognitive: by anchoring the message in accurate content at both the opening and close (Amazeen et al., 2018). In the context of the credibility and trust pathway, the truth sandwich may be particularly effective because it avoids the primary risk of standard prebunking, namely that audiences fixate on the misinformation itself, while retaining the transparency signal that makes prebunking credibility-enhancing in the first place (Ecker et al., 2022). Tulin et al. (2025) suggest that truth sandwiches have a longer-lasting effect, with users showing higher discernment in subsequent messages. Both prebunked posts and truth sandwiches are therefore expected to engage the credibility-to-trust pathway by positioning the RKI as forthright and misinformation-aware, though the truth sandwich may produce stronger credibility gains by mitigating the risk of audience fixation on the misinformation itself.

2.3 Experimental Design

Both studies reported in this paper were implemented using a shared experimental infrastructure and procedural logic, allowing for direct comparability across communication strategies while testing substantively distinct interventions. Specifically, the research consists of two parallel experiments examining (i) uncertainty framing and (ii) prebunking as health communication strategies. Although analytically treated as separate experiments, both studies were embedded within the same custom-built, web-based social media platform, TWON, and followed an identical experimental flow from recruitment through outcome measurement.

Across both experiments, participants were recruited online through Prolific and randomly assigned to experimental conditions prior to interacting with the platform[1]. Participants then completed a short pre-survey capturing demographic characteristics, political orientations, and social media usage patterns (see Table 2.1 for an overview of the overall sample as well as descriptive statistics by experimental component), after which they were directed to an interactive social media feed designed to mimic contemporary platforms in both appearance and functionality.

[1] Each participant could only take part in one of the studies.

Table 2.1: Sample characteristics by condition

Variable	Total (N = 1,066)	Uncertainty (N = 424)	Prebunking (N = 642)
Age, mean (SD)	31.77 (10.05)	31.6 (10.2)	31.9 (9.97)
Political Views			
Political interest, mean (SD)[+]	5.12 (1.36)	5.12 (1.30)	5.12 (1.39)
Ideology, mean (SD)[*]	−0.23 (2.41)	−0.29 (2.36)	−0.18 (2.43)
Gender			
Male	649 (60.9%)	258 (60.8%)	391 (60.9%)
Female	404 (37.9%)	160 (37.7%)	244 (38.0%)
Other	13 (1.2%)	6 (1.4%)	7 (1.1%)
Education			
Upper secondary / A-levels	779 (73.4%)	310 (73.1%)	469 (73.1%)
Applied sciences qualification	141 (13.2%)	60 (14.2%)	81 (12.6%)
Intermediate secondary certificate	97 (9.1%)	42 (9.9%)	55 (8.6%)
Other / No qualification	49 (4.6%)	12 (2.8%)	37 (5.8%)
Employment			
Employed (incl. apprentices, parental leave, etc.)	641 (60.1%)	247 (58.3%)	394 (61.4%)
Student / Pupil	337 (31.6%)	140 (33.0%)	197 (30.7%)
Unemployed	54 (5.1%)	25 (5.9%)	29 (4.5%)
Other	37 (3.5%)	12 (2.8%)	22 (3.4%)
Social Media Use			
Several times daily or more	756 (70.9%)	299 (70.5%)	457 (71.2%)
Once daily	162 (15.2%)	71 (16.7%)	91 (14.2%)
At least once a week	66 (6.2%)	22 (5.2%)	44 (6.9%)
Less frequently or never	82 (7.7%)	32 (7.5%)	50 (7.8%)

[+] Interest ranges from 1 (No interest) to 7 (very interested).

[*] Ideology ranges from −5 (left) to 5 (right).

After completing the pre-survey, participants were directed to the main feed of the TWON platform. Each participant viewed a total of eight posts, presented in random order to approximate the dynamic experience of a real social media feed. Seven of these posts were identical for all participants across both experiments and consisted of content drawn from real-world institutional accounts such as Netflix and Sky Sports, covering a mix of neutral topics including politics, entertainment, and health (see Figure 2.2 for an example). The eighth post was the experimental stimulus, which depended on the participant's assigned condition, and was presented as a post authored by the Robert Koch Institute (in all studies and conditions), lending the health communication content a recognizable and authoritative source identity.



Figure 2.2: Generic post within Studies 1 and 2

In the uncertainty framing experiment (Study 1), participants were exposed to vaccine-related content that varied in how scientific uncertainty was communicated, comparing an uncertainty-framed treatment to a corresponding control condition. In the prebunking experiment (Study 2), participants encountered vaccine-related misinformation or corrective content, comparing a control condition to prebunking interventions designed to build resistance to misinformation, including variants employing truth-sandwich structures. Due to the sensible nature of healthcare communication, all treatment posts were created in conjunction with the Robert Koch Institute[2]. We report on the nature of the treat-

[2] https://www.rki.de/EN/Topics/Infectious-diseases/Immunisation/Information-material/Vaccination-myths/effectively-debunking-misinformation-node.html

ments and the findings of our two studies separately.

Participants were instructed to scroll through the entire feed and read all posts as they would in a typical social media environment. To encourage active engagement, they were asked to like at least two posts and comment on at least three posts, ensuring that they interacted with the content rather than passively viewing it. This procedure allowed the experimental manipulation to be embedded within a realistic, interactive setting, enabling measurement of participants' responses to both the experimental and neutral content under conditions that mimic social media use. Following exposure to the feed, participants of both experiments completed an identical post-survey measuring perceived post credibility and trust in the RKI. Perceived post credibility was assessed using three items asking participants to rate the experimental post on its reliability, accuracy, and trustworthiness, each measured on a seven-point scale ranging from 1 (not at all) to 7 (completely). These three items were combined into a single latent variable in a structural equation model (SEM), specifically, confirmatory factor analysis (CFA). Trust in the RKI was measured using a single item asking participants to rate their overall trust in the Robert Koch Institute, also on a seven-point scale ranging from 1 (no trust at all) to 7 (complete trust). Figure 2.3 gives an overview of the distribution of these variables across conditions and experiments [3].

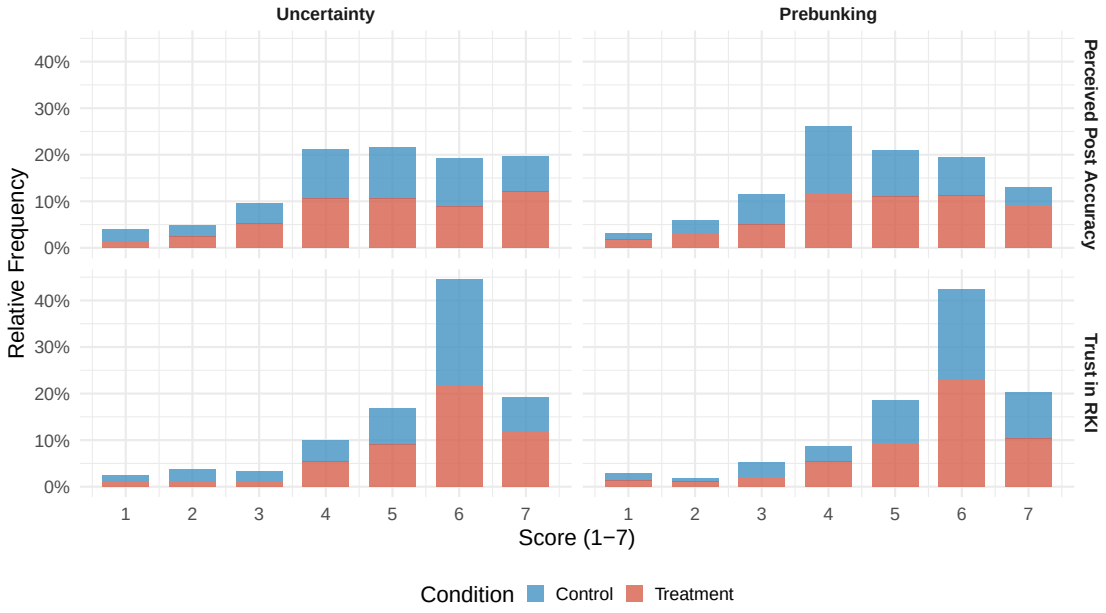


Figure 2.3: Distributions of Trust in RKI and Post accuracy across experiments and conditions

[3] For simplicity, post Accuracy was chosen as a proxy for the latent variable, results are similar for post trustworthiness and reliability.

2.4 Study 1: Uncertainty Framing

This study examines whether explicitly acknowledging scientific uncertainty in a vaccine-related social media post affects perceived post credibility and institutional trust. Participants were randomly assigned to either a treatment or control condition. Both conditions presented the same core health information, namely the 82% effectiveness rate of the Mpox (monkeypox) vaccine, displayed as a prominent statistic in an image attributed to the Robert Koch Institute (see Figure 2.4). In the treatment condition, this was accompanied by a short text that explicitly acknowledged the limits of current scientific knowledge, noting that questions remain about the duration of vaccine protection and how effectiveness may change with new variants, while still concluding with a clear recommendation to vaccinate (see Box 1). In the control condition, participants saw the same image and vaccine statistic without any accompanying qualifying text, isolating the effect of uncertainty acknowledgement from the underlying health information itself.

> More and more people are becoming infected with Mpox (also known as monkeypox). New studies confirm that the vaccination is 82% effective against the disease. Nevertheless, there are still open questions in research, for example how long the protection lasts exactly and how the effectiveness changes with new variants. Vaccination is recommended in order to ensure the best possible protection.

Box 1: Text accompanying the uncertainty framing treatment post, as displayed on the TWON platform.

Demographic characteristics of the uncertainty framing subsample ($N = 424$) are reported in Table 2.1. We model the hypothesized causal chain using confirmatory factor analysis (CFA). Specifically, we test whether exposure to uncertainty framing influences perceived post credibility, a latent variable indicated by three items assessing the reliability, accuracy, and trustworthiness of the post (`PostCredibility`), which in turn predicts trust in the Robert Koch Institute (`TrustRKI`).

The model fit itself is excellent: CFI $= 1.000$, TLI $= 1.009$, RMSEA $= 0.000$ (90% CI $[0.000, 0.067]$), SRMR $= 0.009$, indicating that the structural model fits the data well. The results indicate uncertainty framing had no significant effect on perceived post credibility ($b = 0.213$, $p = .118$), and neither the direct effect on trust in the RKI ($b = 0.097$, $p = .461$) nor the indirect effect via post credibility ($b = 0.077$, $p = .136$) reached significance. The total effect was similarly null ($b = 0.174$, $p = .214$). More notably, the credibility-to-trust pathway itself performed as theoretically expected. Post credibility was a strong and significant predictor of trust in the RKI ($b = 0.360$, $p < .001$), suggesting that the structural

model is well-specified and that the absence of treatment effects reflects the manipulation rather than any flaw in the underlying theoretical pathway. In short, the uncertainty framing did not shift how participants evaluated the post, but had it done so, those credibility perceptions would have translated meaningfully into institutional trust. Figure 2.5 shows the results of the CFA model.

Our findings suggest that uncertainty framing, at least as implemented here, does not influence how audiences evaluate vaccine-related health content. Our interpretation is that the difference between presenting a statistic with and without qualifying language (as our treatment did) is simply insufficient to shift credibility perceptions in an ecologically valid social media environment, where audiences process content quickly and competing posts may dilute the salience of subtle message features - at least for our highly educated sample. A more substantive interpretation, however, is that uncertainty framing has no meaningful effect on credibility perceptions regardless of context. This would be consistent with research suggesting that lay audiences do not consistently reward transparency about scientific limits, particularly when the underlying message, here a clear vaccine recommendation, remains unchanged (Kreps and Kriner, 2020; van der Linden and Leiserowitz, 2019). The credibility-to-trust pathway being internally valid but unactivated by the manipulation points toward the intervention itself as the limiting factor, rather than any flaw in the theoretical model.

2.5 Study 2: Prebunking

This study examines whether prebunking vaccine-related misinformation on a social media post affects perceived post credibility and institutional trust. As in study 1, participants were randomly assigned to one of three conditions: a control condition, a prebunking condition, and a truth sandwich condition, all conditions being equally represented. All three conditions presented content related to the mRNA vaccine, and, in the context of our custom platform, were posted by the RKI account. In the control condition, participants only saw a straightforward informational post without reference to misinformation, shown in Figure 2.6. In both prebunking and truth sandwich conditions, the post also featured a myth surrounding Messenger ribonucleic acid (mRNA) vaccines, as shown in image 2.7. In the truth sandwich condition, the treatment post featured an additional message that further clarified the specific controversies surrounding mRNA vaccines (see Box 2).



Figure 2.4: Uncertainty Framing Stimuli

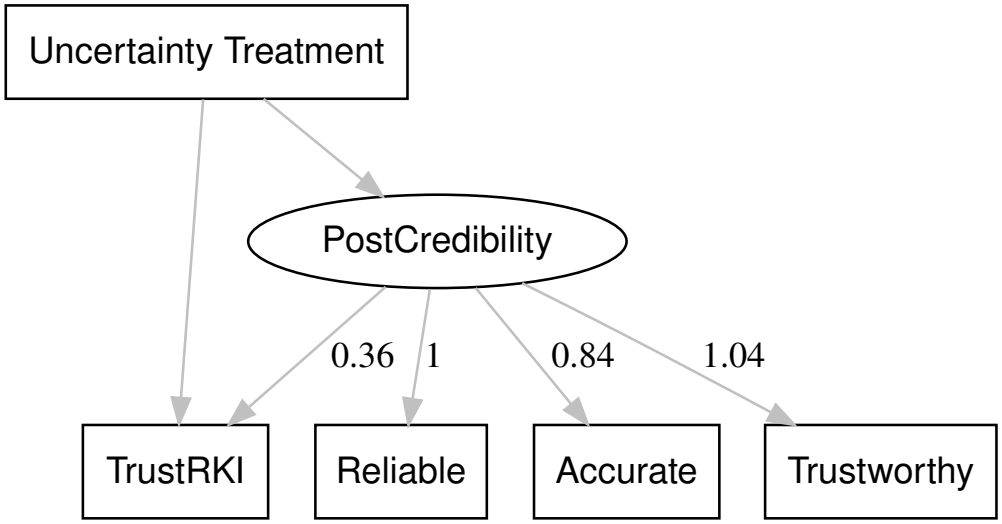


Figure 2.5: Structural Equation Model of the Uncertainty Framing Study

mRNA vaccines are a relatively new technology and have been administered to many millions of individuals within a short time. A certain scepticism as to the effects this might have is therefore quite understandable. It is also understandable that hearing about possible changes to our genetic make-up initially provokes uncertainty. It requires extensive knowledge of biology and cellular function to recognise which assertions are true and which are not, and to understand how mRNA vaccines work and why they are not “gene therapy”.
What is important to know is that mRNA (messenger RNA) occurs naturally in every cell of the human body – in what is called our cytoplasm. Human DNA, by contrast, is always inside the cell’s nucleus. The mRNA from a vaccine cannot reach there. It transports a part of the blueprint of the SARS-Coronavirus-2 exclusively into the cytoplasm – but it cannot penetrate into the nucleus of the human cell.

Box 2: Text accompanying the prebunking treatment post, as displayed on the TWON platform.

Demographic characteristics of the prebunking subsample ($N = 642$) are reported in Table 2.1. As in Study 1, we model the hypothesized causal chain using a structural equation model (SEM), testing whether exposure to either prebunking condition influences perceived post credibility, which in turn predicts trust in the Robert Koch Institute. Both treatment conditions are entered simultaneously into the model, with the control condition serving as the reference category, allowing us to estimate and compare the effects of prebunking alone and the truth sandwich independently.

The model fit is good: CFI $= 0.997$, TLI $= 0.994$, RMSEA $= 0.037$ (90% CI $[0.000, 0.070]$), SRMR $= 0.009$, indicating that the structural model represents the observed data well across a larger and more varied sample than Study 1. Neither prebunking alone ($b = 0.221$, $p = .168$) nor the truth sandwich ($b = 0.125$, $p = .422$) had a significant effect on perceived post credibility, suggesting that neither intervention was sufficient to shift how participants evaluated the message itself. The indirect effects via post credibility were likewise non-significant for both prebunking alone ($b = 0.088$, $p = .171$) and the truth sandwich ($b = 0.050$, $p = .423$), meaning that the hypothesised credibility-to-trust pathway was not activated by either manipulation. The direct effects on trust in the RKI followed a similar pattern, with prebunking alone falling just short of conventional significance ($b = 0.226$, $p = .065$) and the truth sandwich remaining clearly non-significant ($b = 0.140$, $p = .236$). Figure 2.8 visualizes the results of the CFA model.

One result departs partially from this otherwise consistent pattern of null findings. The total effect (in-

direct and direct) of prebunking alone on trust in the RKI did reach significance ($b = 0.314, p = .022$), whereas the total effect of the truth sandwich did not ($b = 0.190, p = .152$). While statistically significant, this effect is practically modest, a standardised coefficient of 0.314 is conventionally characterised as a small-to-medium effect and is unlikely to reflect a meaningful change in underlying attitudes. Moreover, given that the indirect effect through post credibility was non-significant, this total effect is difficult to interpret within the theoretical framework proposed here.

As in Study 1, the credibility-to-trust pathway itself was strongly supported regardless of experimental condition. Post credibility was a robust and substantial predictor of trust in the RKI ($b = 0.397, p < .001$), replicating the structural result across an independent and larger sample. Taken together, the findings from both studies point to a consistent conclusion: the theoretical pathway from credibility to institutional trust is empirically well-founded, but the interventions tested here did not move credibility perceptions sufficiently to activate it.

The results of Study 2 are consistent with those of Study 1 in all essential respects. Across both prebunking conditions, the interventions did not significantly shift perceived post credibility, and the credibility-to-trust pathway was consequently not activated. The truth sandwich, despite its more structured approach to addressing misinformation, did not outperform standard prebunking on any outcome measure. The one partial exception remains the borderline total effect of prebunking alone on trust in the RKI, which reached significance at the conventional threshold but without a corresponding indirect effect through credibility, making it difficult to interpret within the present framework. The replication of the strong credibility-to-trust association ($b = 0.397, p < .001$) across a larger and independent sample nonetheless strengthens confidence in the structural model itself. We discuss the implications of these findings in the following section.

2.6 Discussion

This paper examined whether two theoretically motivated health communication strategies, uncertainty framing and prebunking increase perceived post credibility and institutional trust in an ecologically valid social media environment. Across both experiments, the interventions did not produce significant effects on perceived post credibility, and the hypothesized credibility-to-trust pathway was not activated by either manipulation. In Study 1, uncertainty framing had no significant effect on how participants evaluated the vaccine-related post, nor on their trust in the RKI. In Study 2, neither prebunking alone nor the truth sandwich shifted credibility perceptions sufficiently to activate the pathway, though



Figure 2.6: Prebunking Control



Figure 2.7: Prebunking Treatment

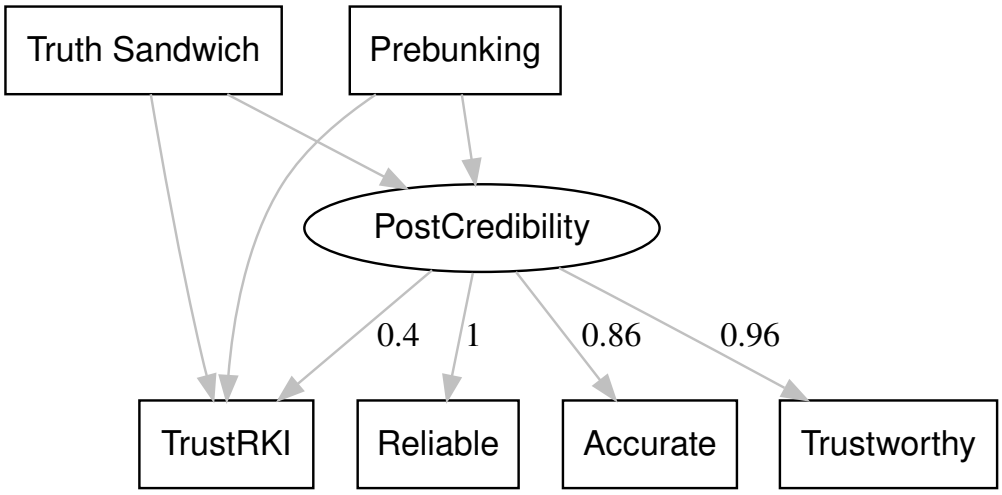


Figure 2.8: Structural Equation Model of the Prebunking Study

the total effect of prebunking alone on institutional trust approached and reached conventional significance. Across both studies, however, the structural pathway from perceived post credibility to institutional trust was strongly and consistently supported, replicating across independent samples and estimation strategies.

The null effect of uncertainty framing is consistent with a line of research suggesting that lay audiences do not reliably reward transparency about scientific limits (Kreps and Kriner, 2020; van der Linden and Leiserowitz, 2019; Roozenbeek et al., 2020). When the underlying recommendation, here a clear encouragement to vaccinate, remains unchanged, the addition of qualifying language about the limits of scientific knowledge may simply not register as a meaningful signal. Participants exposed to the uncertainty-framed post may have processed the core message, the 82% effectiveness statistic and the vaccination recommendation, without attending carefully to the epistemic qualifications surrounding it. This interpretation is consistent with dual-process accounts of message processing (Petty and Cacioppo, 1986), which suggest that peripheral cues such as hedging language are often discounted under conditions of moderate involvement.

The absence of significant effects in Study 2 is more surprising given the substantial body of evidence supporting prebunking as an effective intervention (Lewandowsky and van der Linden, 2021; van der Linden et al., 2017; Roozenbeek and Van der Linden, 2019). The failure of prebunking to shift credibility perceptions in this setting suggests that the mechanism through which inoculation effects typically operate, namely the motivated and cognitive processing of corrective information, may require more sustained engagement than a single social media post embedded within a busy feed can provide (Basol et al., 2021). The truth sandwich, despite its theoretically motivated structural advantage in minimizing the memorial footprint of misinformation (Amazeen et al., 2018), similarly failed to outperform standard prebunking on any outcome measure. This null comparison is consistent with recent evidence suggesting that truth sandwich formats do not reliably produce immediate effects on belief correction or credibility evaluations despite having long-term discernment benefits (Tulin et al., 2025). The absence of notable effects in both conditions suggests that the structural refinement offered by the truth sandwich did not confer additional credibility benefits in this context, at least not via the credibility pathway tested here.

A consistent thread running through both sets of null results is the role of the experimental setting itself. Traditional studies on health messaging have typically relied on surveys or vignette-based experiments that expose participants to health content in controlled isolation (Kreps and Kriner, 2020; van der Lin-

den and Leiserowitz, 2019; Nan and Madden, 2012). While such approaches provide considerable experimental control, they may fail to capture the dynamic and socially mediated context in which health information is actually consumed online (Lewandowsky and van der Linden, 2021; Guess and Lyons, 2020). The present studies were conducted within a custom-built, interactive social media platform designed to approximate realistic browsing conditions, with participants scrolling through a full feed of competing content before encountering the experimental stimulus. This ecological validity may itself explain the null results: interventions that succeed in focused laboratory settings may lose their potency when they must compete for attention alongside entertainment content, political news, and other health posts. Simply put, in an environment where all posts compete for attention, a nudge to uncertainty or an inoculation might not captivate users' attention (Pennycook et al., 2020).

This methodological contribution is arguably as important as the substantive findings. The consistent failure of theoretically well-motivated interventions to produce effects in a naturalistic setting raises questions about the external validity of the existing prebunking and uncertainty communication literature, much of which is based on isolated message exposure paradigms (Basol et al., 2021; Roozenbeek and Van der Linden, 2019). At the same time, the strong internal validity of the credibility-to-trust pathway across both studies suggests that the theoretical model is sound. The challenge for future research is therefore not to revise the underlying theory but to identify intervention designs that are capable of activating the credibility pathway under the noisy, competitive conditions of real information environments. This may require longer exposure, repeated inoculation, or more visually salient message formats that are better suited to the affordances of social media platforms (Roozenbeek et al., 2020). Understanding how to effectively promote scientific trust in naturalistic information environments remains an urgent challenge for health communication research.

Chapter 3

Nudging civility on OSN Using LLMs

Abstract

Online social networks have become a central arena for political and social discourse, yet interactions on these platforms are frequently characterized by incivil and toxic interactions. While disagreement is a normal and required feature of democratic debate, research suggests that disrespectful communication discourages users from engaging in political discussions and may negatively affect both participants and the broader audience exposed to such interactions. In response, previous interventions have attempted to improve online discourse through behavioral nudges and interface design changes, though their effectiveness has often been limited. This study examines whether AI-mediated paraphrasing interventions can reduce uncivil expression while preserving substantive disagreement in online political discussions. Using an experimental setting that simulates social media interactions, we analyse how AI-generated paraphrases influence the tone of conversations and assess their effects not only on the direct participants of a debate but also on external observers who encounter these exchanges. The findings provide insights into the potential of AI-assisted communication tools to foster healthier online discourse.

3.1 Introduction

Online social networks have become a central arena for political and social discourse (Shah et al., 2005; Wojcieszak and Mutz, 2009). However, a growing body of evidence indicates that politically oriented interactions on social media are frequently characterized by persistent incivility and toxic exchanges (e.g., Mamakos and Finkel, 2023). Although everyday political conversations are seldom idealized forms of public deliberation (Fishkin, 2009), empirical research suggests that it is the tone of online debates, rather than disagreement itself, that most often prompts users to disengage from political discussion (Sydnor, 2019; Bor and Petersen, 2022). Such dynamics are damaging not only to the direct participants of these exchanges but also to the broader set of stakeholders involved, including the audience of users who encounter these interactions without participating in them.

In response to the prevalence and consequences of toxic political discourse, prior efforts to promote healthier online interactions have frequently relied on behavioural nudges and interface interventions aimed at encouraging civility or reflection. While such approaches can yield modest improvements in specific contexts (Mathew et al., 2019; Chetty and Alathur, 2018), their overall effectiveness has often been shown to be limited in comparison to more intensive, interpersonal interventions such as door-to-door conversations (Broockman and Kalla, 2016). Recent advances in artificial intelligence, and generative models in particular, have opened new opportunities for social science by enabling scalable, adaptive, and context-sensitive interventions in human communication (Bail, 2024). Several empirical studies have demonstrated that AI-mediated conversational interventions can reduce toxic language while fostering healthier and more engaging political debates online (Argyle et al., 2023; Combs et al., 2023; Tessler et al., 2024).

Importantly, such interventions may shape not only the experiences of direct participants but also those of the broader audience of users who encounter these exchanges without actively engaging, for whom observed norms of discourse can influence perceptions of political discussion, willingness to participate, and broader democratic attitudes. Building on this motivation, the current study asks: How can AI-mediated paraphrasing interventions reduce uncivil expression while preserving substantive disagreement, and what are their effects on both participants and the external audience of observers in online political discussions?

3.1.1 Toxicity in OSN

Online incivility is often rooted in a combination of psychological, structural, and political factors. Its presence across OSNs is well-documented (Mamakos and Finkel, 2023; Siegel, 2020) and has increased markedly in recent years. At the individual level, affective polarisation and perceived out-group threat intensify hostile communication, as users increasingly treat political opponents not merely as ideological adversaries but as morally illegitimate (Iyengar et al., 2019). These dynamics are compounded by platform architectures that reward outrage and emotional arousal; algorithmic amplification and anonymity have each been shown to lower the thresholds for hostile expression (Rösner and Krämer, 2016; Brady et al., 2017; Yu et al., 2024). Partisan sorting further accelerates this process, as ideologically homogeneous networks reinforce in-group identity and normalise antagonistic discourse directed at out-groups (Bail et al., 2018). Understanding the sources of online hostility is, however, complicated by the heterogeneity of the phenomenon itself. Rossini (2022) argues that "incivility" and "intolerance" must be treated as distinct phenomena: while the former often accompanies substantive political argument and may carry some deliberative value, the latter directly attacks the legitimacy of out-groups and poses a graver threat to democratic norms. This conceptual distinction carries significant consequences for intervention design, since moderation strategies calibrated only to surface-level incivility risk suppress legitimate, if impassioned, political speech, while failing to adequately target the more corrosive forms of intolerance that erode the foundations of democratic participation.

Regardless of where on this spectrum hostile speech falls, its aggregate effects on the quality of democratic life are well-documented. Synthesising a substantial experimental and meta-analytic literature, Van't Riet and Van Stekelenburg (2022) find consistent evidence that exposure to political incivility reduces political trust and dampens the willingness to participate in public debate. Notably, this implies that it is the tone of online political exchange, rather than disagreement per se, that most powerfully suppresses citizens' motivation to engage (Bor and Petersen, 2022).

3.1.2 Common Approaches

Responses to online incivility have historically been centred around two poles: content moderation and counter speech. First, content moderation, encompassing account bans, post removal, and algorithmic demotion, has attracted sustained criticism for its reactive character, inconsistent application, and free-expression implications (Rossini, 2022; Gillespie, 2018). The empirical record on its effectiveness is mixed: while removal of egregiously harmful content has been shown to reduce its immediate spread (Jhaver et al., 2019), evidence that it durably changes user behaviour or suppresses incivility

at scale remains limited, with some studies suggesting that deplatformed users simply migrate to less regulated spaces (Horta Ribeiro et al., 2021).

Second, counter speech, which responds to harmful content with factual corrections, empathetic dialogue, or inclusive narratives, has been positioned as a less restrictive alternative (Cepollaro et al., 2023). Under favourable conditions, it can moderate both the speaker's behaviour and the broader audience (Hangartner et al., 2021). Counter speech, however, is highly context-dependent and resource-intensive. Both approaches are predominantly reactive and externally imposed, a limitation that has prompted growing scholarly interest in automated, pre-emptive moderation tools, and more recently in the application of large language models to detect and intervene in uncivil exchanges before they escalate.

3.1.3 Generative AI as an intervention

Within social sciences, generative AI has moved beyond its initial role as a classifier of harmful content to become part of the infrastructure through which human behaviour online is actively shaped (Bail, 2024). A growing literature suggests that LLMs are capable not only of detecting incivility but of intervening before it occurs. Conversational agents have been shown to durably shift attitudes and reduce hostile beliefs (Costello et al., 2024). Likewise, micro-targeted LLM-generated messages have demonstrated meaningful persuasive effects in political contexts (Tappin et al., 2023). This logic has been extended to online moderation directly, with studies finding that real-time LLM interventions can reduce the hostility of political exchanges and facilitate debates across ideological lines (Combs et al., 2023; Argyle et al., 2023).

What remains unclear, however, is how such interventions perform under the noisier, more anonymous conditions characteristic of real social media environments and whether their effects extend beyond the immediate exchange to the participants who join the conversation later. The aforementioned studies testing LLM moderation in dyadic settings implicitly model online interaction as a bounded exchange between two parties (Argyle et al., 2023; Combs et al., 2023). This dyadic design often does not accurately represent social media, where conversations are public, asynchronous, and continuously joined (or just read) by new participants with no exposure to earlier interventions. The contribution of our study is to address this by moving toward a standard social media setting, where multiple users participate asynchronously, not necessarily reacting to one another. We aim to go beyond asking whether treated comments become more civil in isolation, and ask instead whether LLM interventions translate into healthier discourse at the conversational level. That is, whether the moderation of individual

comments creates a more civil environment for subsequent users joining the conversation.

3.2 Methods

3.2.1 Experimental Design and Platform

The study employed a between-subjects experimental design with two conditions: (1) an LLM-treatment condition in which participants received optional AI-generated rewrite suggestions for their posts, and (2) a control condition in which participants posted freely without such prompts. Ethical approval was provided by the University of Amsterdam. The experiment was conducted on a custom-built simulated social network platform, the TWON ("Twin of Online Networks") platform, designed explicitly to replicate the core interaction dynamics of real-world OSNs. The platform was hosted externally and integrated with Prolific for participant recruitment and compensation. The TWON environment allowed participants to self-select into discussions on topics of personal interest, to observe and respond to contributions from multiple other users. This design choice was made to approximate the conditions of real-world political discourse on OSNs, where individuals are not assigned discussion partners but rather choose which conversations to join and how to engage.

The study recruited a sample of $N = 574$ German-speaking adults (aged 18+) residing in Germany via Prolific All participants provided informed consent prior to participation, and all interactions took place in German. Participants were compensated at a rate of 10 pounds per hour, with the study session lasting approximately 10 minutes. Prolific's pre-screening tools were used to ensure participants met the residency and language criteria. The experiment was pre-registered and lasted from January 3rd to 23rd, 2026. The preregistration, the experiment materials and scripts for the analysis can be found here `https://osf.io/gmbex/overview?view_only=d305d1ee704b4f44bee4365704c64466`.

3.2.2 Procedure

Upon enrolling in the study, participants first completed a brief pre-survey that collected demographic information and self-reported political orientation. Participants were then introduced to the TWON platform and instructed that they would be participating in a series of short online discussions on contemporary political and social topics. Participants were presented with a menu of 10 topics, spanning issues chosen to reflect a range of political salience and affective polarisation. This self-selection mechanism was intended to mirror the voluntary nature of participation in real OSN discussions, where users gravitate towards content aligned with their interests or concerns.

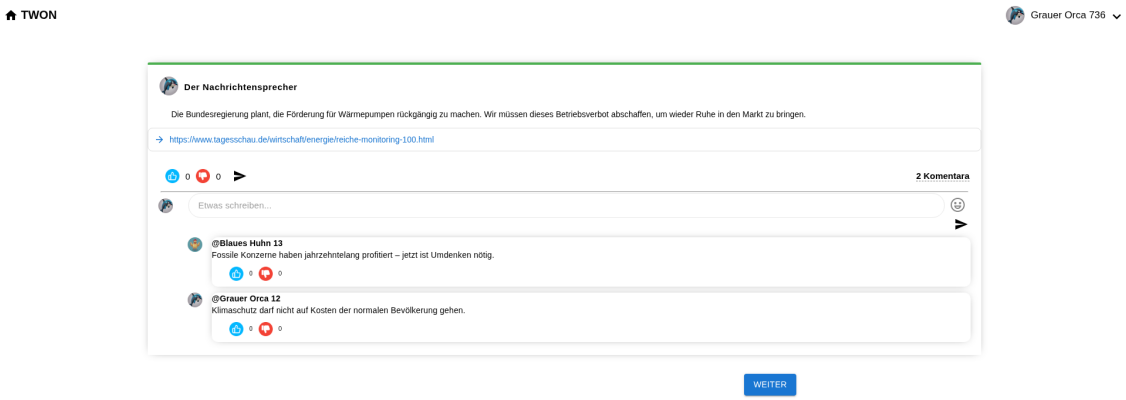


Figure 3.1: Example of a post in the TWON Feed

Participants were then invited to contribute their own response. Conversations were allowed to reach a maximum depth of 8 comments. Once a thread reached this maximum depth, the system initialised a new discussion thread on the same topic (With identical seed-comments and posts), ensuring a continuous flow of fresh conversational contexts. For each selected topic, participants viewed the following (see Figure 3.1 for an example):

1. A short initial post, authored by the research team. For half (5) of the topics, a link to a news article was provided in this initial post. For all other topics, no additional information beyond the text of the post was given.

2. 2 to 3 topic-dependent inflammatory comments authored by the research team. These comments were implemented to foster a toxic environment. These pre-set comments were indistinguishable from participant-made comments[1].

3. Any existing comments contributed by earlier participants.

In the LLM-treatment condition, the intervention occurred at the moment of message submission. After a participant typed their comment and clicked the submit button, a pop-up window appeared displaying an AI-generated rewrite of their message. This rewrite was generated in real time by OpenAI's `GPT-5` model regardless of the initial toxicity of the comment. The LLM model was instructed to preserve the semantic content and stance of the original message while reducing incivility, hostility, or inflammatory language. After being prompted with the LLM-paraphrase participants were given three options: (1) post the AI-suggested rewrite, (2) manually edit the rewrite before posting, or (3) discard the rewrite and post their original message unchanged. The intervention was fully opt-in and invisible to other par-

[1] The first participant to join a discussion would only see these comments.

ticipants. In the control condition, participants encountered no such intervention. Upon clicking *post*, their message was posted immediately without modification. After completing their participation in all selected discussion threads, participants answered a brief post-survey that assessed their subjective experience of the platform.

The primary dependent variable was comment toxicity (later aggregated at the discussion level), operationalised using the Perspective API developed by Google. Perspective API provides automated scores across multiple dimensions of harmful speech, including overall toxicity, severe toxicity, insult, profanity, threat, and identity-based attack. Each dimension is scored on a continuous scale from 0 (no toxicity detected) to 1 (high toxicity). The API has been widely adopted in computational social science research as a reliable and validated measure of online incivility (Dixon et al., 2018).

3.3 Analysis and Results

The final analytical sample consisted of 574 participants who completed the full experimental protocol. None of the participants failed both attention checks in our survey; 37 of them failed a single attention check. We do not exclude those participants from our analysis. Table 3.1 presents the demographic composition of the sample, overall and by experimental condition. Gender was male-balanced, with 37.108% female and 61.1450% male participants. Age ranged from 18 to 73 years, with a mean of 29.683 years (SD = 9.158). Politically, participants self-reported their ideological orientation on a 10-point scale (-5 = far left, 5 = far right). Politically, the sample was skewed toward the left with an average score of -0.9686 (SD = 2.026), with 0 being the neutral centre of the 10-point scale. Random assignment to experimental conditions was successful, as confirmed by the absence of statistically significant differences in demographic characteristics, educational attainment, or political orientation between the LLM-treatment and control groups (all $p < 0.01$).

Participants were allowed to select discussion topics based on their personal interests. On average, participants commented 6.068 times, with a minimum requirement of 4. The most frequently selected topics were the *Deutschland Ticket*[2] ticket with 523 comments and *Cannabis legalisation* with 465 comments. The least popular were *Waterpumps' Governmental Subsidy* with 188 comments and *Gendered language* with 230 comments. Among participants assigned to the LLM-treatment condition ($n = 452$, representing 78.746% of the total sample for a total of 2721 treated comments for 762 untreated comments). Within these treated users, paraphrase suggestions were given for 2026 comments, no modi-

[2] The Deutschland ticket topic focused on proposed plans to raise its price to 63 euros (from 58 euros).

Table 3.1: Sample Demographics by Treatment Response

	Control	Treatment
N (participants)	122	452
Comment rate	6.246	6.020
Age Mean (SD)	30.492 (9.702)	29.465 (9.004)
Political Leaning* Mean (SD)	-0.992 (2.002)	-0.962 (2.034)
Gender		
Female Proportion (%)	36.066	37.389
Male Proportion (%)	62.295	60.841

*Political leaning is measured on a scale from -5 (left-leaning) to 5 (right-leaning), with 0 indicating a neutral position.

fication being warranted for 695 of the comments. Of the suggested modifications, 1884 were posted as-is or after manual editing. [3]. We found no statistically significant differences in age, gender or political leaning in the propensity to accept or reject the paraphrases.

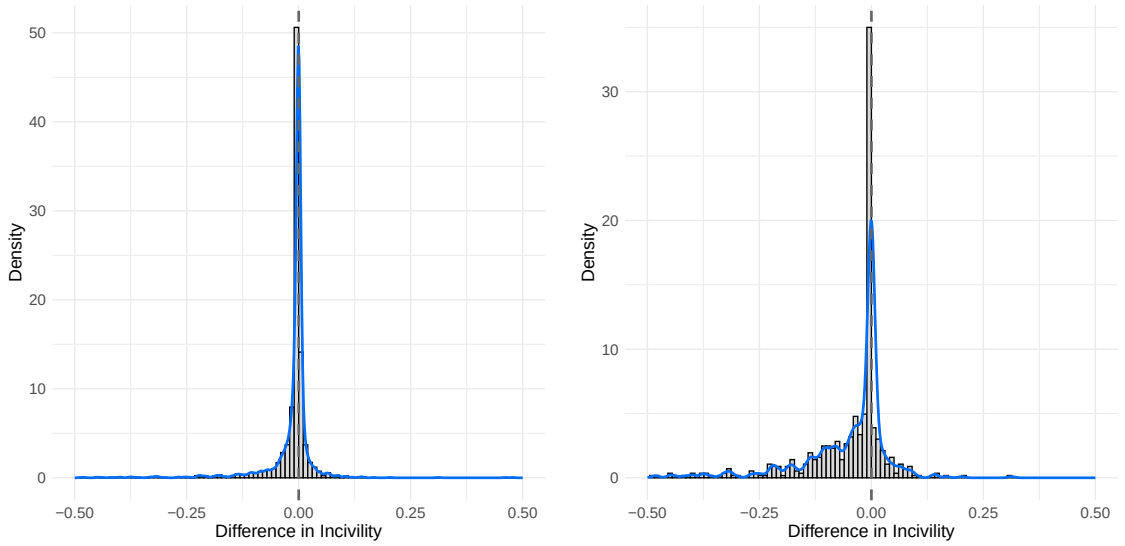
3.3.1 LLMs and Toxicity reduction

Message Level

To assess the immediate effectiveness of the LLM paraphrasing intervention, we first examine toxicity at the comment level, i.e., the toxicity of individual comments as ultimately posted to the platform. This analysis can also be regarded as a manipulation check verifying whether the AI-generated rewrites successfully reduce the toxicity of participants' comments. The degree to which participants chose to adopt less toxic formulations and the extent to which those formulations actually reduced measured toxicity is a finding in its own right.

For each message, we computed toxicity scores across six dimensions using the Perspective API: `toxicity`, `severe toxicity`, `insult`, `profanity`, `threat`, and `identity attack`. Although we collected data on all dimensions, our primary outcome measure is `incivility`[4], here defined as the highest score across any of the six specific dimensions for a given message. This composite metric captures the most salient form of incivility present in each comment.

[3] Only 12 suggestions were further edited (as opposed to posted as-is).
[4] We use the `incivility` metric as a mean to capture the entirety of the perspective API results, not per the definition of **?**



(a) Toxicity Difference across all comments

(b) Toxicity Difference across originally unicivil comments ($\geq .1$)

Figure 3.2: Differences in toxicity between original and paraphrased comments

We distinguish between two versions of each message produced by participants in the treatment condition: the original comment (the text initially typed by the participant before any LLM intervention) and the posted comment (the final text that appeared on the platform, which could be either the original message if the participant declined the rewrite, the LLM-generated paraphrase if accepted as-is, or a manually edited version of the paraphrase). For participants in the control condition, the original and posted comments are identical by design, as they received no opportunity to revise their text. Figure 3.2 presents the distribution of toxicity change across all messages in the treatment condition. Negative values indicate that the posted message was less toxic than the original, while positive values indicate increased toxicity. The distribution is left-skewed, indicating that the majority of LLM-assisted rewrites reduced toxicity relative to participants' initial drafts. However, it is worth noting that most participants were already relatively civil in their original contributions: the median original `incivility` score was 0.029. While persistent, this indicates that toxicity and incivility were a minority behaviour in the sample's context.

Table 3.2 presents formal paired-sample t-tests comparing original and posted toxicity scores across all six Perspective API dimensions and our aggregating metric for the full treatment sample. Overall toxicity declined from a mean of 0.073 in original messages to 0.060 in posted messages ($p < 0.001$). This indicates that there is a steady decrease in toxicity when LLMs were used to moderate user comments before they are posted. The small nature of the difference can be explained by the polite nature of most comments. When looking at more abrasive comments (set by a minimum score of `incivility`), we no-

tice that the differences in all metrics become more pronounced. We note that the `threat` dimension is never statistically significant across toxicity levels. This is likely due to the low occurrence of threats in the dataset.

Taken together, these findings establish that the LLM paraphrasing intervention achieved its goal: when participants accepted AI-generated rewrites, the resulting posted messages were measurably less toxic than their original drafts. The effect was present across most toxicity dimensions and was most pronounced for comments that were initially toxic. This message-level analysis does not yet address the core theoretical question motivating the study, i.e. whether reducing the toxicity of individual comments has downstream effects on the quality of the broader conversation.

Conversation Level

Beyond message-level reductions, we examine whether the LLM intervention affects toxicity at the conversation level, that is, whether reducing the toxicity of one comment influences the tone of subsequent contributions from other participants. This addresses the core theoretical question: do civility interventions have spillover effects on the broader discussion, including among users who did not receive the intervention themselves?

We model conversation-level toxicity using two dependent variables: (1) `FollowUpToxicity`, the toxicity of the immediately next comment in the thread, and (2) `DownstreamToxicity`, the mean toxicity of all subsequent comments. We control for user-level random effects, comment depth (log-transformed number of prior comments), and discussion topic (10 fixed categories). Given the right-skewed distribution of toxicity scores, we employ a log-log specification [5] shown in Equations 3.1 and 3.2, with weakly informative priors specified in Equations 3.3 to 3.6. The model was estimated using Bayesian hierarchical regression in Stan with 4 chains of 2,000 iterations each (1,000 warm-up). All chains converged, for all parameters R-hat < 1.01, and $N_{eff} \geq 400$. This indicates that the Markov chains mixed well and produced stable, reliable posterior estimates.

$$\begin{aligned}\log(\text{ConversationToxicity}_i) =& \beta_0 + \beta_1 \log(\text{AcceptedTreatments}_i) + \\ & \beta_2 \log(\text{ConversationDepth}_i) + \\ & \boldsymbol{\beta_3}^\top \text{Topic}_i + \gamma_{\text{user}[i]} + \varepsilon_i \end{aligned} \tag{3.1}$$

[5] In a log-log model, coefficients are interpreted as elasticities: a 1% change in the independent variable corresponds to a β% change in the dependent variable.

Table 3.2: Comparisons of Toxicity Scores Between Original and Posted Comments

Attribute	Original	Posted	p_{adj}
All Messages (n = 2,698)			
Toxicity	0.070	0.058	0.000***
Severe Toxicity	0.004	0.003	0.000***
Insult	0.043	0.034	0.000***
Profanity	0.027	0.019	0.000***
Threat	0.012	0.011	0.092
Identity Attack	0.022	0.019	0.000***
Incivility	0.073	0.060	0.000***
High Incivility (Original or Posted $\geq$ 0.10, n = 604)			
Toxicity	0.215	0.176	0.000***
Severe Toxicity	0.013	0.008	0.000***
Insult	0.144	0.109	0.000***
Profanity	0.077	0.045	0.000***
Threat	0.028	0.024	0.266
Identity Attack	0.078	0.064	0.000***
Incivility	0.225	0.182	0.000***

Note. Paired t-tests comparing original (pre-intervention) and posted (post-intervention) toxicity scores for messages in the treatment condition. p_{adj} values reflect Bonferroni correction for multiple comparisons. *** $p_{adj} < .001$.
High incivility messages are classified based on the `incivility` metric. If either the posted (i.e. treated) or the original comment exceeds the specified threshold, it is considered in the analysis.

$$\text{ConversationToxicity}_i \in \{\text{FollowUpToxicity}_i,\ \text{DownstreamToxicity}_i\} \tag{3.2}$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. We specify weakly informative priors:

$$\beta_k \sim \mathcal{N}(0, 1), \quad k = 0, 1, 2, 3 \tag{3.3}$$

$$\sigma \sim \mathcal{N}^+(0, 1) \tag{3.4}$$

$$\sigma_\gamma \sim \mathcal{N}^+(0, 1) \tag{3.5}$$

$$\gamma_j \sim \mathcal{N}(0, \sigma_\gamma^2), \quad j = 1, \ldots, N_{\text{users}} \tag{3.6}$$

Figure 3.3 and Table 3.3 present the posterior distributions for the coefficients of both models. For both dependent variables, the treatment shows a statistically significant, albeit small, effect. Specifically, a 1% increase in accepted treatments corresponds to a 0.026% reduction in `FollowUpToxicity` (95% CI [-0.036, -0.016]) and a 0.017% reduction in `DownstreamToxicity` (95% CI [-0.024, -0.009]). Our results suggest that the LLM intervention's effects extend beyond the directly treated messages. By reducing the toxicity of individual comments, the intervention lowers the ambient hostility of the conversation, which in turn reduces the likelihood that subsequent participants—including those in the control condition—will adopt an uncivil tone. This spillover effect suggests that civility interventions need not be universally applied to produce conversation-wide benefits.

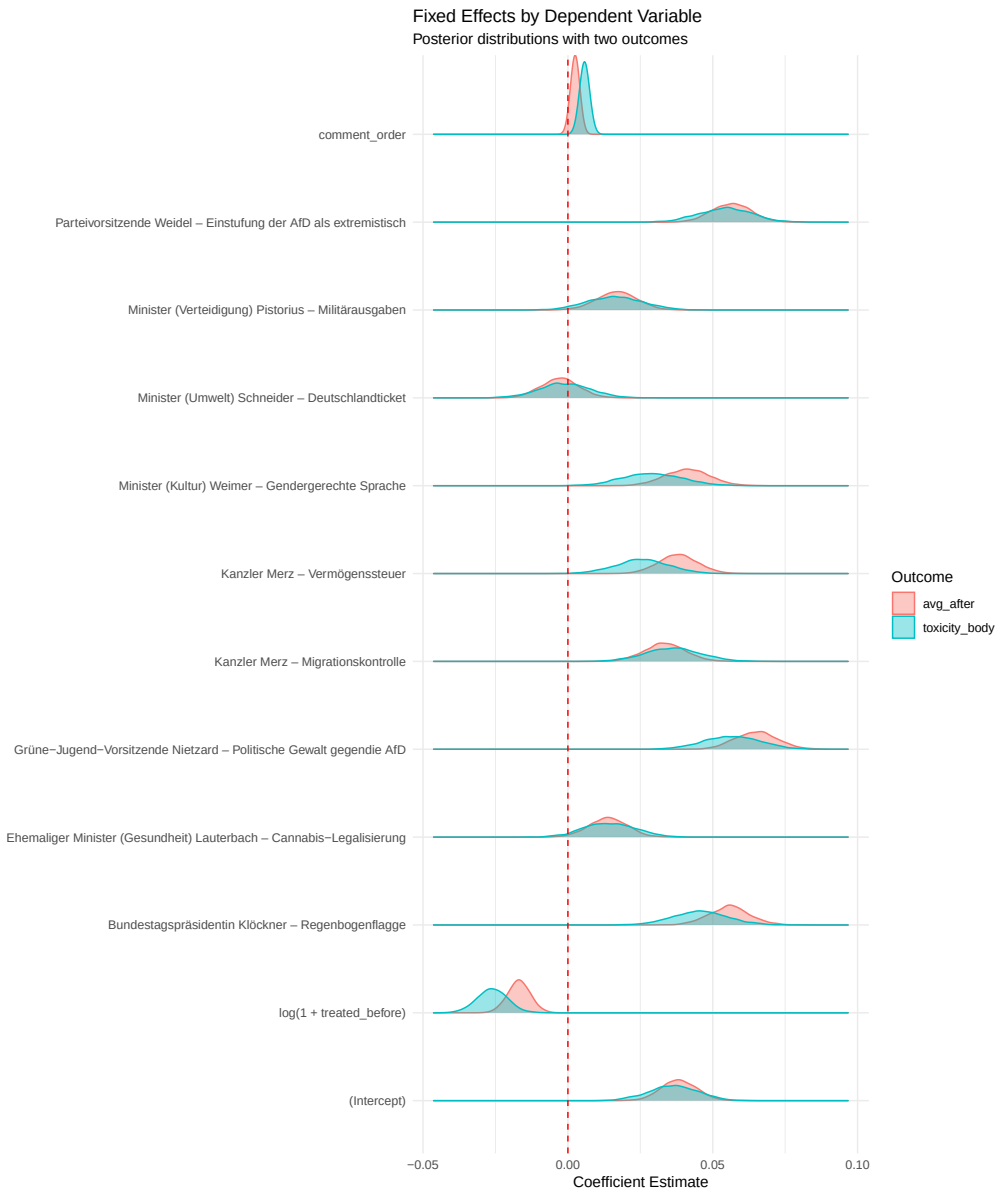


Figure 3.3: Posterior distributions and 95% credible intervals for key regression coefficients.

Table 3.3: Hierarchical Regression Results of Conversation-Level Toxicity

	DownstreamToxicity				**FollowUpToxicity**			
Parameter	Estimate	*SD*	95% CI	$\hat{R}$	Estimate	*SD*	95% CI	$\hat{R}$
Fixed Effects								
Intercept	0.038	0.006	[0.027, 0.050]	1.00	0.036	0.008	[0.020, 0.053]	1.00
log(1 + Accepted Treatments)	-0.017	0.004	[-0.024, -0.009]	1.00	-0.026	0.005	[-0.036, -0.016]	1.00
log(Conversation Depth)	0.056	0.007	[0.042, 0.070]	1.00	0.046	0.009	[0.027, 0.064]	1.00
Model Fit								
Marginal R^2	0.173	0.017	[0.141, 0.206]	1.00	0.080	0.012	[0.057, 0.104]	1.00
Conditional R^2	0.185	0.018	[0.149, 0.221]	1.00	0.140	0.018	[0.107, 0.176]	1.00
Prop. Var. (Treatment)	0.023	0.010	[0.007, 0.046]	1.00	0.030	0.011	[0.011, 0.056]	1.00

Note. Bayesian hierarchical models estimated using Stan with 4 chains of 2,000 iterations (1,000 warm-up). All $\hat{R} < 1.01$ indicating convergence (N = 3483). *SD* = posterior standard deviation. Credible intervals are 95% highest posterior density intervals. Random effects for users and discussion topics not shown.

3.4 Discussion & Conclusion

This study tested whether LLM-assisted paraphrasing can reduce toxicity in online political discourse. We deployed the intervention in a simulated social network environment that allowed asynchronous discussions with self-selected participation. Specifically, we examine whether civility interventions affect discourse quality for subsequent participants who join conversations later—including those who never interact with the intervention themselves.

Threads containing LLM-moderated comments exhibited lower toxicity in subsequent contributions, including from control participants and from users who joined after the treated participant had left. Prior experimental work on AI-mediated discourse has largely relied on dyadic designs where all participants engage with the intervention (Argyle et al., 2023; Combs et al., 2023). Our findings suggest that treated comments can influence the tone of entire threads without direct player interactions.

The pattern is consistent with a norm-setting mechanism. When early comments in a thread are less hostile, later participants adopt a more measured tone. This relates to work showing that tone, rather than disagreement, drives political disengagement (Bor and Petersen, 2022; Sydnor, 2019), and to research on emotional contagion in networks (Brady et al., 2017). The effect persists after the treated user has left, as the less toxic comment continues to shape subsequent discourse. Although the effect sizes are modest, two important considerations are worth noting. First, they likely reflect the inherent unpredictability and contextual variability of online discourse. Second, even small reductions in toxicity at the individual message level may accumulate, such that the overall impact across a discussion thread could exceed the immediate per-message effect. Whether such compounding dynamics emerge in longer conversations, however, remains an open empirical question.

From a practical perspective, pre-publication paraphrasing differs from post-hoc removal, which may displace users rather than change behaviour (Horta Ribeiro et al., 2021), and from counter-speech (Cepollaro et al., 2023) or offline interventions (Broockman and Kalla, 2016), which do not scale without sustained labour. The opt-in design preserves user agency while producing measurable effects. As Bail (2024) notes, generative AI has become part of the infrastructure shaping online behaviour.

The findings indicate that LLM-based interventions can generate spillover effects for users who do not directly engage with the system. Operating in asynchronous settings, the intervention affects conversational norms without requiring universal coverage. The potential for compounding effects suggests this class of interventions warrants further investigation in naturalistic settings.

Part II

Simulation studies

Overview of Simulation Studies

In this second part of the report, we investigate whether fully automated simulations can be effectively used to complement and extent the results found in the case studies of Part I. Whereas traditional experiments are suited to capture how a particular treatment affects opinions, attitudes, or behaviours of users, simulation-based studies are able to represent the collective dynamics that emerge when many users interact within a shared information environment. This makes simulations a valuable tool for studying systematic phenomena such as the spread of misinformation, the formation of echo chambers, and the amplification of polarising content. We present two short simulation studies in Part II. In the first simulation study, we simulate the experimental design of the second case study. Specifically, we investigate whether truth sandwich investigations can contribute to a democratically desirable conversation behavior. In a second simulation study, we synthetically reproduce the real-world behaviour of Serbian users towards the Russo-Ukrainian conflict. The underlying goal for these simulations is to be used as an alternative to the first case study presented in Chapter 1.

Chapter 4

Simulation Debate Quality

4.1 Introduction

Social media has become a primary source for health information for large parts of the population (cites in comment). The tone and content of those discussions shape how individuals think about vaccination, risk, and institutional guidance (Sanders et al., 2023; Zhao and Zhang, 2017; Roozenbeek et al., 2020). Conversations dominated by incivility or misinformation can normalise vaccine hesitancy and erode trust in public health authorities in the long term (Jiang, 2024; Viviani and Pasi, 2017). At the same time, discussions that are rational, engaged, and informative may, on the one hand, reinforce accurate beliefs (Jiang, 2024; Tsugawa and Ohsaki, 2017; Jin et al., 2016), however, on the other hand, are less likely to be promoted (Trilling et al., 2017; Hornik et al., 2015). Understanding whether an institutional actor like the RKI can shift discussions in this direction, therefore, has direct implications for how public health communication should be designed and deployed.

This chapter builds directly on the experimental study reported in Section 2, which tested whether prebunking or a truth sandwich can reduce beliefs in vaccine misinformation, as well as debate quality measures developed in an early stage of the TWON project. Our previous case study found no statistically conclusive direct effects on perceived credibility of the prebunking or truth sandwich intervention, suggesting that these health communication strategies do not convincingly change individual minds in a social media setting.

In this chapter, we address the question of whether posts from a renowned institution such as the RKI has indirect positive effects on the online debate surrounding vaccination: Do RKI posts change the

tone of the discussions that form around them? Unlike the study in chapter 2, answering this requires observations from many users interacting repeatedly over time, and documenting shifting opinions induced by constant exposure - Conditions that a controlled lab experiment cannot reproduce accurately. Therefore, using the TWON architecture, we model the effect of an RKI account injecting multiple truth sandwich interventions into a synthetic social network, and track how markers of debate quality – incivility, rationality, and interactivity – evolve across the network as a result. The simulation used here was developed in collaboration with researchers at the University of Trier (Work Package 3[1]). What follows is a brief, non-technical overview of its main components. For a full technical description, we refer the reader to Schwager et al. (2025).

4.2 Simulation Detail

To examine the effect of health-related institutional communication on the quality of online discourse, we conducted two parallel simulations using the Conditioned Comment Prediction (CCP) framework introduced by Schwager et al. (2025). Each simulation populated a simulated social media environment with $1,000$ LLM-based user agents. The two conditions were identical in all respects except one: in the treatment condition, an additional agent modelled on the Robert Koch Institute was present and active throughout.

User agents were instantiated using open-weight language models[2] and conditioned on their individual interaction histories rather than explicit persona descriptions. This implicit conditioning approach, in which each agent's behaviour is grounded in up to 15 prior stimulus-response pairs drawn from authentic digital traces across English, German, and Luxembourgish social media corpora, was shown by Schwager et al. (2025) to consistently outperform biography-based alternatives in terms of both lexical and semantic fidelity to real user behaviour. At each of the 288 simulation steps, agents were exposed to posts selected by a recommender system operating on semantic similarity, (computed as the cosine distance between post embeddings), approximating the filter bubble dynamics observed on real platforms.

The RKI agent was modelled with a disproportionately large network, reflecting its high real-world follower count. At every step, it posted a truth sandwich, a structured correction format that acknowledges a misconception, refutes it with evidence, and reinforces accurate information, drawn from ma-

[1] Another Work package in the TWON project
[2] Llama 3.1- 8B-Instruct

terials originally published on the official RKI website [3]. Across both conditions, we tracked three debate quality metrics, *Incivility*, *Rationality*, and *Interactivity*, as developed by Stolwijk et al. (2026), respectively capturing the hostility of exchanges, the degree to which contributions are evidence-based and on-topic, and the extent to which users engage with one another rather than posting in isolation. Stolwijk et al. (2026) considerably expands on the validity and use cases of these metrics.

4.2.1 Results and Discussion

Table 4.1 shows the debate quality metrics for both experimental conditions. The first column summarises the simulation without and the second column with an active RKI account. Across all debate metrics, we see that the introduction of the active RKI account produced small improvements in the desired direction of online debate quality. *Incivility* declined marginally from 32.40% to 32.05%, a reduction of 0.35 percentage points. *Rationality* increased from 27.36% to 28.73%, a gain of 1.37 percentage points. *Interactivity* showed the smallest absolute change, rising from 5.91% to 5.96%. Thus, even though a simple truth sandwich post may not produce the envisioned impact on increasing trust in science, the post nonetheless has a desired effect on how the online discussion evolves and thus can affect how users engage with the discussion. Though none of these changes are large, the *Rationality* gain of 1.37 percentage points is the most notable, amounting to a 5% improvement over the baseline.

Table 4.1: Debate quality metrics by experimental condition

Metric	No RKI Account	With RKI Account	Direction
Incivility	32.40%	32.05%	↓ (desired)
Rationality	27.36%	28.73%	↑ (desired)
Interactivity	5.91%	5.96%	↑ (desired)
Diversity	77.15%	78.56%	↑ (desired)

These results suggest that institutional health communication can exert a small but consistent positive influence on the quality of online discourse. The uniform direction of the effects across all three metrics is encouraging and unlikely to be coincidental, though the effect sizes remain modest throughout. This pattern is consistent with the experimental findings reported in Chapter 2, where RKI truth-sandwich posts also failed to produce large shifts in individual credibility assessments. Taken together, both strands of evidence point to the same conclusion: a single institutional voice, however well-resourced,

[3] https://www.rki.de/EN/Topics/Infectious-diseases/Immunisation/Information-material/Vaccination-myths/effectively-debunking-misinformation-node.html

cannot substantially reshape the texture of online debate on its own.

This has practical implications for public health communication strategy. Rather than relying on high-volume posting from a central account, interventions may need to engage a broader ecosystem of trusted messengers, target users at structurally influential network positions, or combine communication efforts with platform-level design changes. Future simulation work could test these alternatives directly by varying the number of pro-vaccine accounts (here only formally represented by the RKI), their network centrality, or the frequency and framing of their posts.

Chapter 5

Towards simulating Serbian debates

5.1 Introduction

One of the core aims of TWON was to not only study social networks in frequently-studied Western, often English-language, contexts. From a technical point of view, in low-resourced languages, it may be more difficult to build well-functioning LLM-based agents; and also content-wise, other political systems than the often-studied Anglo-centric two-party-systems may lead to different debate dynamics on social media. Therefore, it is crucial to broaden the linguistic and geographical scope.

In our work on Serbian discussions on the Russian war against Ukraine described in Section 1, we gained first experiences in empirically studying the subject. In this chapter, we now describe work towards simulating such debates.

5.2 A first LLM-based simulation

In a first step, we explored the feasibility of creating LLM-based agents in Serbian[1]. This relatively small-scale experiment simulated a synthetic social media environment, populated by 100 agents. The agents could write posts, but also like or dislike existing posts, or decide to follow another agent. In particular, the agents were prompted in such a way that they resemble 100 different persona's, based on characteristics typically also used in survey research on social media use. These include next to sociodemographic characteristics, psychological traits, political attitudes, and social-media-based features such

[1] The full study is reported in Bojić et al. (2025)

as connectivity or engagement levels.

The different agents then played in multiple rounds, allowing to assess multiple outcome metrics, in particular engagement levels, affective polarization, and the evolution of follower networks. Importantly, the system allowed to test multiple different platform mechanics, in particular different recommender systems and interventions such as boosting specific categories of content, such as news vs. entertainment.

This first study demonstrated the feasibility of the work flow. In particular, the results indicated that the platform mechanics indeed mattered and let to different outcome metrics.

5.3 Calibrating LLM-agents on Serbian survey respondents

While the first study, which is briefly summarized in the previous section and reported in full in Bojić et al. (2025), led to encouraging results, we decided that a more systematic calibration of the agents was necessary[2].

To this end, we fielded a survey among a large, representative sample of Serbian internet users between November 21 and November 28, 2025. It consisted of two parts: the collection of respondent characteristics, and their reaction to (fictive) social media posts. Ultimately, our goal was to create agents that based on the information from part 1, would be able to replicate the behaviour in part 2.

In part 1, we collected basic sociodemographic information (age, gender, education, employment status, and region of residence) and attitudes that may be relevant for the topics that are discussed in the social-media posts. These include institutional trust (toward the army, healthcare system, media, science, government, police, church, NGOs, and trade unions), political orientation and geopolitical attitudes (toward the Russia-Ukraine conflict, EU integration, Trump, and China), attitudes toward conspiracy theories and ecology, personality traits, leisure time preferences (e.g., reading, gaming), music genre preferences (e.g., classical, pop), and engagement in daily activities such as cooking, shopping, volunteering, interior design. Additionally, we asked about their use of ChatGPT.

In part 2, we showed participants in total $N_{posts} = 56$ different posts, deliberately splitting them between posts that are related and posts that are unrelated to the attitudes measured before. Participants could then react to these posts, by liking, disliking, commenting, or sharing – or by not reacting at all.

[2] Based on a manuscript in preparation: Bojić, L., Dinić, B., Ilić, V., Mevorah, V., Trilling, D. & Rettinger, A.: Toward Universal Social Media Simulation with LLM-Based AI Agents: Benchmarking Reaction Accuracy Using 120K+ Persona Versions of 1511 Humans

Liking occured the most (43.9% across all posts), followed by no reaction at all (30.3%), disliking (19.7%), while commenting (6.8%) and sharing (2.8%) happened least. This is quite in line with what one would expect, except maybe that in real-world settings, one would of expect no-reaction to be even more prevalent. Yet, this is quite explainable given the setting, in which participants are explicitly asked to do do something.

For each of the $N_{respondents} = 1511$ valid and complete responses, we then proceeded to construct an LLM agent. In particular, we were interested in which level of persona specificity would result in agents that mimic human behaviour sufficiently well.

We compared the performance of $N_{models} = 27$ different LLMs $\times$ $N_{prompts} = 3$ different versions of persona specificity $\times$ $N_{validrespondents} = 1511$ personas, hence creating in total $N_{agents} = 121,500$ unique agent-persona combinations.

Our results indicate that, all things considered, the choice of the LLM is the by far largest determinant of accuracy. Including more additional information from the variables collected in part 1 of the survey only increased the accuracy of the predictions marginally. The full prompt yielded the highest accuracy (71.5%), followed by a prompt only including value-related questions judgments(71.2%), while the demographics prompt showed the lowest accuracy (69.6%). But again, the differences are smaller than one might expect, given that the full prompt included information directly relevant to the behaviour (namely, attitudes towards the very topics the posts were about).

The following steps are now merging these agents into a simulation framework like described in Section 5.2.

Conclusion

This report was organised around three primary goals. The first was to address substantive questions posed by our societal partners – the Robert Koch Institute (RKI) and the Slovenian Press Agency (STA) – on topics of direct public relevance. The second was to develop and validate the TWON platform as a credible experimental infrastructure for studying social media dynamics under controlled yet realistic conditions. The last objective was to use the case studies as a source of training data for the simulations.

In response to STA's interest in the role of platform mechanics in disinformation engagement, Study 1 found that recommender systems, whether reinforcing or opposing, did not systematically shape how Serbian participants engaged with distorted content about the Russo-Ukrainian war. With regards to RKI's interests in vaccine communication, Study 2 found that neither prebunking, truth-sandwiches, nor uncertainty framing significantly shifted perceived post credibility or institutional trust when embedded within a realistic social media feed. Both studies underscored the importance of ecological validity in communication research. The effects of recommender systems and prebunking-style interventions already carry mixed support in the academic literature Guess et al. (2019); Tulin et al. (2025); Lewandowsky and van der Linden (2021); Guess et al. (2018), and our case studies suggest that the ecological constraints of a realistic social media environment make such effects even harder to detect. This is a meaningful finding in its own right, despite its negative nature. It suggests that effect sizes observed under controlled isolation may not straightforwardly translate to the noisy, competitive conditions under which people actually consume information online.

Each of the three studies contributed to platform improvements. With study 1, we made substantial improvements to the backend infrastructure and scalability of the platform, allowing multiple users to be active on TWON at the same time. Study 2 validated that TWON is working as intended. It demonstrates its applicability to research questions that have hardly been tested on social media. Having gained confidence in the internal and external validity, study 3 provides us with an opportunity to test features that

are not currently present on social media. This is hardly researched in social science, as setting such an experiment up without a platform such as TWON is difficult.

The case studies were originally designed to provide the simulations with data for their initial parameters. Over the course of the project, however, the state of the art shifted considerably toward LLM-based approaches for agent creation, substantially changing the required training data. Rather than serving as a starting point for the simulations, the case studies became complementary. Whereas our case studies excelled at isolating the causal effects of interventions on individual behaviour, they are poorly suited to studying systematic, network-level dynamics. The simulations of this study are meant to address this limitation precisely. In particular, we show that LLM-based agents can effectively and simply approximate the behaviour of OSN users. We also simulate the setting of our second case study, finding that informative posts made by the RKI improve the quality of the debates surrounding the issue.

Bibliography

Esma Aïmeur, Sabrine Amri, and Gilles Brassard. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30, 2023. 16

Michelle A Amazeen, Emily Thorson, Ashley Muddiman, and Lucas Graves. Correcting political and consumer misperceptions: The effectiveness and effects of rating scale versus contextual correction formats. *Journalism & Mass Communication Quarterly*, 95(1):28–48, 2018. 25, 29, 39

Lisa P Argyle, Christopher A Bail, Ethan C Busby, Joshua R Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate. Leveraging ai for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences*, 120 (41):e2311627120, 2023. 42, 44, 55

Christopher A Bail. Can generative ai improve social science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121, 2024. 42, 44, 55

Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221, 2018. 16, 17, 23, 43

Mert Basol, Jon Roozenbeek, and Sander van der Linden. Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of Cognition*, 4(1):2, 2021. 28, 39, 40

Ljubiša Bojić, Velibor Ilić, Veljko Prodanović, and Vuk Vuković. An Agent-Based Simulation of Politicized Topics Using Large Language Models: Algorithmic Personalization and Polarization on Social Media.

Chinese Political Science Review, December 2025. ISSN 2365-4244, 2365-4252. doi: 10.1007/s41111-025-00326-x. 62, 63

Alexander Bor and Michael Bang Petersen. The psychology of online political hostility: A comprehensive, cross-national test of the mismatch hypothesis. *American political science review*, 116(1):1–18, 2022. 42, 43, 55

William J Brady, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318, 2017. 43, 55

Elena Broda and Jesper Strömbäck. Misinformation, disinformation, and fake news: lessons from an interdisciplinary, systematic literature review. *Annals of the International Communication Association*, 48(2):139–166, 2024. 16

Urie Bronfenbrenner. Toward an experimental ecology of human development. *American psychologist*, 32(7):513, 1977. 22

David Broockman and Joshua Kalla. Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science*, 352(6282):220–224, 2016. 42, 55

David V. Budescu, H. H. Por, and Steven B. Broomell. Communicating uncertainty in science and public policy: The role of numbers, visuals, and explanations. *Psychological Science*, 32(3):353–372, 2021. 25, 27

Bianca Cepollaro, Maxime Lepoutre, and Robert Mark Simpson. Counterspeech. *Philosophy Compass*, 18(1):e12890, 2023. 44, 55

Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H Chi. Top-k off-policy correction for a reinforce recommender system. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 456–464, 2019. 23

Naganna Chetty and Sreejith Alathur. Hate speech review in the context of online social networks. *Aggression and violent behavior*, 40:108–118, 2018. 42

Aidan Combs, Graham Tierney, Brian Guay, Friedolin Merhout, Christopher A Bail, D Sunshine Hillygus, and Alexander Volfovsky. Reducing political polarization in the united states with a mobile chat platform. *Nature human behaviour*, 7(9):1454–1461, 2023. 42, 44, 55

Thomas H Costello, Gordon Pennycook, and David G Rand. Durably reducing conspiracy beliefs through dialogues with ai. *Science*, 385(6714):eadq1814, 2024. 44

Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pages 191–198, 2016. 16, 23

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, 2018. 47

Ullrich KH Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K Fazio, Nadia Brashier, Panayiota Kendeou, Emily K Vraga, and Michelle A Amazeen. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1):13–29, 2022. 29

James Fishkin. *When the people speak: Deliberative democracy and public consultation*. Oup Oxford, 2009. 42

Seth Flaxman, Sharad Goel, and Justin M Rao. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, 80(S1):298–320, 2016. 17

Tarleton Gillespie. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018. 43

Andrew Guess, Brendan Nyhan, Benjamin Lyons, and Jason Reifler. Avoiding the echo chamber about echo chambers. *Knight Foundation*, 2(1):1–25, 2018. 22, 65

Andrew Guess, Jonathan Nagler, and Joshua Tucker. Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science advances*, 5(1):eaau4586, 2019. 17, 22, 65

Andrew M Guess and Benjamin A Lyons. Misinformation, disinformation, and online propaganda. *Social media and democracy: The state of the field, prospects for reform*, 10:10–33, 2020. 17, 25, 27, 28, 40

Abel Gustafson and Ronald E Rice. The effects of uncertainty frames in three science communication topics. *Science Communication*, 41(6):679–706, 2019. 27

Abel Gustafson and Ronald E Rice. A review of the effects of uncertainty in public science communication. *Public Understanding of Science*, 29(6):614–633, 2020. 27

Dominik Hangartner, Gloria Gennaro, Sary Alasiri, Nicholas Bahrich, Alexandra Bornhoft, Joseph

Boucher, Buket Buse Demirci, Laurenz Derksen, Aldo Hall, Matthias Jochum, et al. Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, 118(50):e2116310118, 2021. 44

Jacob Hornik, Rinat Shaanan Satchi, Ludovica Cesareo, and Alberto Pastore. Information dissemination via electronic word-of-mouth: good news travels fast, bad news travels faster! *Computers in Human Behavior*, 45:273–280, 2015. 58

Manoel Horta Ribeiro, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Gianluca Stringhini, Emiliano De Cristofaro, and Robert West. Do platform migrations compromise content moderation? evidence from r/the_donald and r/incels. *Proceedings of the ACM on Human-Computer Interaction*, 5 (CSCW2):1–24, 2021. 44, 55

Shanto Iyengar, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood. The origins and consequences of affective polarization in the united states. *Annual review of political science*, 22 (1):129–146, 2019. 43

Shagun Jhaver, Amy Bruckman, and Eric Gilbert. Does transparency in moderation really matter? user behavior after content removal explanations on reddit. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–27, 2019. 43

Shaohai Jiang. Does social media promote or hinder health learning? the roles of media attention, information discussion, information elaboration, and information seeking experience. *Mass Communication and Society*, 27(4):627–652, 2024. 58

Jiahua Jin, Xiangbin Yan, Yijun Li, and Yumei Li. How users adopt healthcare information: an empirical study of an online q&a community. *International journal of medical informatics*, 86:91–103, 2016. 58

Sarah E. Kreps and Douglas L. Kriner. Model uncertainty, political contestation, and public trust in science: Evidence from the covid-19 pandemic. *Science Advances*, 6(43):eabd4563, 2020. 25, 26, 27, 34, 39

Sophie Lecheler and Jana Laura Egelhofer. Disinformation, misinformation, and fake news: Understanding the supply side. *Knowledge resistance in high-choice information environments*, pages 69–87, 2022. 16

Stephan Lewandowsky and Sander van der Linden. Countering misinformation and fake news through

inoculation and prebunking. *European Review of Social Psychology*, 32(2):348–384, 2021. doi: 10.1080/10463283.2021.1876983. URL https://doi.org/10.1080/10463283.2021.1876983. 25, 26, 28, 39, 40, 65

Michalis Mamakos and Eli J Finkel. The social media discourse of engaged partisans is toxic even when politics are irrelevant. *PNAS nexus*, 2(10):pgad325, 2023. 42, 43

Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. Thou shalt not hate: Countering online hate speech. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 369–380, 2019. 42

W. J. McGuire. The effectiveness of supportive and refutational defenses in immunizing and restoring beliefs against persuasion. *Journal of Abnormal and Social Psychology*, 63(2):326–332, 1961. 28

Simon Münker, Nils Schwager, and Achim Rettinger. Don't Trust Generative Agents to Mimic Communication on Social Networks Unless You Benchmarked their Empirical Realism, 2025. arXiv:2506.21974 [cs]. 11

Simon Münker, Nils Schwager, Kai Kugler, Michael Heseltine, and Achim Rettinger. Next Reply Prediction X Dataset: Linguistic Discrepancies in Naively Generated Content, 2026. URL http://arxiv.org/abs/2602.19177. 11

Xiaoli Nan and Kelly Madden. Hpv vaccine information in the blogosphere: how positive and negative blogs influence vaccine-related risk perceptions, attitudes, and behavioral intentions. *Health communication*, 27(8):829–836, 2012. 25, 40

C Thi Nguyen. Echo chambers and epistemic bubbles. *Episteme*, 17(2):141–161, 2020. 22

Brendan Nyhan and Jason Reifler. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330, 2010. 17

Eli Pariser. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin, 2011. 16, 17, 22

Gordon Pennycook and David G. Rand. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, 117(27): 15545–15548, 2020. 28

Gordon Pennycook and David G Rand. The psychology of fake news. *Trends in cognitive sciences*, 25(5): 388–402, 2021. 17, 22

Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu, and David G Rand. Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science*, 31(7):770–780, 2020. 28, 40

Richard E Petty and John T Cacioppo. The elaboration likelihood model of persuasion. In *Advances in experimental social psychology*, volume 19, pages 123–205. Elsevier, 1986. 39

Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio AF Almeida, and Wagner Meira Jr. Auditing radicalization pathways on youtube. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 131–141, 2020. 16, 22, 23

Jon Roozenbeek and Sander Van der Linden. Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, 5(1):65, 2019. 25, 28, 39, 40

Jon Roozenbeek, Claudia R Schneider, Sarah Dryhurst, John Kerr, Alexandra LJ Freeman, Gabriel Recchia, Anne Marthe Van Der Bles, and Sander Van Der Linden. Susceptibility to misinformation about covid-19 around the world. *Royal Society open science*, 7(10), 2020. 25, 26, 27, 28, 39, 40, 58

Leonie Rösner and Nicole C Krämer. Verbal venting in the social web: Effects of anonymity and group norms on aggressive language use in online comments. *Social Media+ Society*, 2(3): 2056305116664220, 2016. 43

Patrícia Rossini. Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research*, 49(3):399–425, 2022. 43

Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *science*, 311(5762):854–856, 2006. 22

Catherine Sanders, Kristin Gibson, Allison R Byrd, Tatevik Markosyan, and Alexa J Lamm. Exploring the social media health information seeking patterns of rural residents to provide communication strategies for extension. *Journal of Applied Communications*, 107(4):1, 2023. 58

Nils Schwager, Simon Münker, Alistair Plum, and Achim Rettinger. Towards simulating social media users with LLMs: Evaluating the operational validity of conditioned comment prediction. In *Proceed-*

ings of the 15th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA), 2025. 59

Nils Schwager, Simon Münker, Alistair Plum, and Achim Rettinger. Towards Simulating Social Media Users with LLMs: Evaluating the Operational Validity of Conditioned Comment Prediction, 2026. URL `http://arxiv.org/abs/2602.22752`. 11

Dhavan V Shah, Jaeho Cho, William P Eveland Jr, and Nojin Kwak. Information and expression in a digital age: Modeling internet effects on civic participation. *Communication research*, 32(5):531–565, 2005. 42

Alexandra A Siegel. Online hate speech. *Social media and democracy: The state of the field, prospects for reform*, pages 56–88, 2020. 43

Sjoerd B Stolwijk, Mark Boukes, Wang Ngai Yeung, Yufang Liao, Simon Münker, Anne C Kroon, and Damian Trilling. Can we use automated approaches to measure the quality of online political discussion? how to (not) measure interactivity, diversity, rationality, and incivility in online comments to the news. *Communication Methods and Measures*, 20(1):1–25, 2026. 60

Cass R Sunstein. Republic: Divided democracy in the age of social media. 2018. 16, 22

Emily Sydnor. *Disrespectful democracy: The psychology of political incivility*. Columbia University Press, 2019. 42, 55

Ben M Tappin, Chloe Wittenberg, Luke B Hewitt, Adam J Berinsky, and David G Rand. Quantifying the potential persuasive returns to political microtargeting. *Proceedings of the National Academy of Sciences*, 120(25):e2216261120, 2023. 44

Michael Henry Tessler, Michiel A Bakker, Daniel Jarrett, Hannah Sheahan, Martin J Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C Parkes, et al. Ai can help humans find common ground in democratic deliberation. *Science*, 386(6719):eadq2852, 2024. 42

Damian Trilling, Petro Tolochko, and Björn Burscher. From newsworthiness to shareworthiness: How to predict news sharing based on article characteristics. *Journalism & mass communication quarterly*, 94(1):38–60, 2017. 58

Sho Tsugawa and Hiroyuki Ohsaki. On the relation between message sentiment and its virality on social media. *Social network analysis and mining*, 7(1):19, 2017. 58

Marina Tulin, Michael Hameleers, Claes De Vreese, Michaël Opgenhaffen, and Ferre Wouters. Beyond belief correction: Effects of the truth sandwich on perceptions of fact-checkers and verification intentions. *Journalism Practice*, 19(11):2576–2595, 2025. 25, 29, 39, 65

Sander van der Linden and Anthony Leiserowitz. Communicating uncertainty about climate change. *Nature Climate Change*, 9(9):686–693, 2019. 25, 26, 34, 39

Sander van der Linden, Anthony Leiserowitz, Seth Rosenthal, and Edward Maibach. Inoculating the public against misinformation about climate change. *Global Challenges*, 1(2):1600008, 2017. 25, 28, 39

Jonathan Van't Riet and Aart Van Stekelenburg. The effects of political incivility on political trust and political participation: A meta-analysis of experimental research. *Human Communication Research*, 48(2):203–229, 2022. 43

Marco Viviani and Gabriella Pasi. Credibility in social media: opinions, news, and health information—a survey. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 7(5):e1209, 2017. 58

Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *science*, 359 (6380):1146–1151, 2018. 17

Magdalena E Wojcieszak and Diana C Mutz. Online groups and political discourse: Do online discussion spaces facilitate exposure to political disagreement? *Journal of communication*, 59(1):40–56, 2009. 42

Xudong Yu, Magdalena Wojcieszak, and Andreu Casas. Partisanship on social media: In-party love among american politicians, greater engagement with out-party hate among ordinary users. *Political Behavior*, 46(2):799–824, 2024. 43

Yuehua Zhao and Jin Zhang. Consumer health information seeking in social media: a literature review. *Health Information & Libraries Journal*, 34(4):268–283, 2017. 58

twon

twin of online social networks



Contact us

Damian Trilling

Project Coordinator

+31 62 782 7904

d.c.trilling@uva.nl

University of Amsterdam
Postbus 15791
1001 NG Amsterdam

Funded by
the European Union