Conference on Harmful Online Communication (CHOC)
16.-17.11.2023

EXTENDED PROGRAMME

Cologne
GESIS – Leibniz Institute for the Social Sciences
Unter Sachsenhausen 6-8, 50667 Cologne
Germany

& online
Link will be provided after registration

This event is funded by the
Fritz Thyssen Stiftung
Programme Overview

Thursday, 16 November 2023
09:30 – 09:45 Welcome and Introduction
09:45 – 10:45 Opening Panel Discussion
10:45 – 11:00 Coffee break
11:00 – 12:30 Input session 1
12:30 – 14:00 Lunch break
14:00 – 15:30 Input session 2
15:30 – 16:00 Refreshments
16:00 – 17:00 Discussion
17:00 – 18:00 Poster Mini Talks
20:00 Social event (details to be confirmed)

Friday, 17 November 2023
09:30 – 10:30 Input Session 3
10:30 – 11:00 Coffee Break
11:00 – 12.30 Input Session 4
12:30 – 14:00 Lunch break
14:00 – 15:30 Work session
15:30 – 16:00 Closing remarks
Extended Programme

Thursday, 16 November 2023

09:30 – 09:45   Welcome and Introduction
Katrin Weller, GESIS
Pascal Siegers, GESIS
Indira Sen, University of Konstanz
Christina Dahn, GESIS

09:45 – 10:45   Opening Panel Discussion
Harmful Online Communication: societal impact and the role of platform governance
Moderation: Katrin Weller, GESIS

Elena Jung, modus | zad, Centre for Applied Research on Deradicalisation
Paloma Viejo Otero, Center for Media, Communication and Information Research (ZeMKI), University of Bremen
Paul Röttger, Bocconi University

The aim of this first panel is to shift the focus from an academic perspective on Harmful Online Communication and individual research results to a broader view on the interactions between Harmful Online Communication, social media platforms and society.

11:00-12:30   Input session 1
Perspectives on harmful online content: Hate and dehumanization
Moderation: tba

Understanding online threats to politicians
Isabelle van der Vegt, Utrecht University
Abstract: The rising trend of online abuse and threats directed at politicians raises safety concerns and underscores the impact this phenomenon may have on democracy as a whole. The current project linguistically analyses abuse and threats in tweets directed at political party leaders in the Netherlands throughout the entire year of 2022. Results show marked gender differences, with female ethnic minority politicians receiving the highest levels of threats. To conclude, practical implications of this study will be discussed, in addition to future avenues for increasing our understanding of online threats to politicians.

Extreme speech and mis/disinformation: A view from the Global South
Iginio Gagliardone, University of the Witwatersrand
Abstract: Both hate speech and mis/disinformation are normatively loaded concepts. Their analysis is increasingly important to understand transformations in digital ecosystems, but also tends to impose frameworks developed in centres of knowledge in the Global North onto individuals and communities who may share different worldviews (e.g., in terms of the distinctions between rumors and false information, or humour and vitriol). Building on research conducted on conspiracy theories in South Africa and Nigeria, this paper examines what different communities “do” with hate and mis/disinformation, highlighting the situatedness
of extreme speech, and proposing alternative normative frameworks for its interpretation.

**Title tba**  
**Homa Hosseinmardi**, University of Pennsylvania  
Abstract: tba

**The computational social science of conspiracy theory communities: from social factors to moderation strategies**  
**Mattia Samory**, Sapienza University of Rome  
Abstract: Online communities enable a plurality of experiences and expressions. One unintended consequence is that fringe ideas like conspiracy theories become easier to encounter than in the past. Similarly, it becomes easier for the holders of such ideas to convene with like-minded individuals and to form dedicated online communities. This talk will introduce research investigating communities devoted to discussing conspiracy theories, once rare and resistant to academic inquiry. Taking a computational social science approach, the talk will discuss the social forces underpinning how individuals encounter conspiracy theories, join dedicated communities, and become increasingly engaged with them. This perspective will outline the challenges and opportunities around managing the boundary between mainstream and conspiracy communities.

14:00-15:30  
**Input session 2**  
**Perspectives on harmful online content: Linking hate and disinformation / polarization**  
Moderation: Veronika Batzdorfer, GESIS

**Anti-vaccine rabbit hole leads to political representation: the case of Twitter in Japan**  
**Tetsuro Kobayashi**, Waseda University  
Abstract: Anti-vaccine attitudes pose a threat to public health by impeding the development of herd immunity. This study, using Japanese Twitter data, revealed that (a) anti-vaxxers are characterized by high political interest, (b) persistent anti-vaxxers were more ideologically left-leaning and had stronger ties to existing political parties, and (c) pandemic-induced new anti-vaxxers displayed low political engagement but a greater affinity for conspiracy theories, spirituality, naturalism, and alternative health practices, which served as gateways to anti-vaccination views. Furthermore, those who turned anti-vaccine after the pandemic also showed increasing support for the emerging anti-vaccine party, leading to their representation in national politics.

**Factuality in the age of large pre-trained language models**  
**Isabelle Augenstein**, University of Copenhagen  
Abstract: Natural language processing is currently experiencing a golden age, thanks to the emergence of chatbots powered by large pre-trained language models (LLMs), able to produce fluent and coherent responses to user input. This has resulted in a wealth of possibilities and enabled new downstream NLP applications. However, powerful as they might seem at a first glance, LLMs are opaque, and produce hallucinations, i.e. factually incorrect output, if used as is. In this talk, I will briefly discuss their limitations of LLMs. I will then present examples of how to reveal their inner workings, and how to test their outputs for factuality.
Digital media and democracy: what is changing globally and how to measure it
Philipp Lorenz-Spreen, Max-Planck-Institute Berlin

Abstract: Information and communication technology has undergone dramatic developments over the past two decades. Increased peer-to-peer connectivity has led to more self-organised public discourse, but it has also given researchers new tools to quantify precisely this systemic shift. Detailed and longitudinal data from social media allow us to measure and model their network structures and dynamics. However, to get a holistic and global picture, a recent systematic literature review has provided us with a number of dimensions of political behaviour that appear to be influenced by the use of digital media. Our findings show that, while the directions within each dimension are mostly clear, they are distributed differently globally and the mechanisms by which these dimensions are linked are still unknown. Understanding these better is crucial for civil society in democracies worldwide, and I will conclude with a methodological outlook on how we can empirically investigate these missing links in the future.

Insecure LLMs generate and augment hate. They shouldn't.
Leon Derczynski, ITU Copenhagen & University of Washington

Abstract: Dealing with human-origin hate already presents a significant challenge. However, large language models, either alone or as human augmentations, scale up text production, and with this comes the risk of harmful text. This talk examines the notion of LLM security, with concrete methods for assessing and texturing the risk of harm presented by LLMs and tools for determining how secure an LLM is.

16:00-17:00
Discussion
Data access options and their influences on the quality of studying harmful online communication
Moderation: Pascal Siegers, GESIS

17:00-18:00
Poster Mini Talks
Perspectives on harmful online content
Moderation: Maria Zens, GESIS
See appendix

20:00
Social event (details to be confirmed)
Friday, 17 November 2023

09:30-10:30  
**Input session 3**

**Perspectives on platform and country dimensions of harmful online content**  
Moderation: Gabriella Lapesa, GESIS

**Profiling Hate Speech Spreaders & Personalizing Counter Narratives against Hate Speech**  
Ilia Markov, Vrije Universiteit Amsterdam

Abstract: In order to effectively reach hate speech authors with counter speech, it is important to understand who the creators of hateful content are and to personalize counter narratives based on their demographic profiles. In this talk, I will provide insights into which kind of people are more likely to post hateful online content and discuss various strategies for effectively generating personalized counter narratives against hate speech.

**Countering Harmful Online Communication in Brazil: Predicting Fine-Grained Factuality of News and Offensive Context of Social Media Comments**  
Francielle Vargas, University of São Paulo

Abstract: The constant increase of harmful online communication (e.g. hate speech and fake news) around the world has become an urgent global problem. Nowadays, most existing automated fact-checking address article-level analysis of news. Nevertheless, news credibility and fact-checking systems at scale require accurate prediction, since each document comprises multiple sentences, which may contain factual information, bias, and misinformation. For hate speech detection, there are also a wide range of challenges including offensiveness and hate speech definitions, missing contextual information and scarce consideration of their social bias. In this talk, we will discuss our advances towards addressing these limitations in order to counter misinformation and hateful comments in Brazil. In this regard, we proposed accurate annotation schemas and developed different data resources for sentence-level factuality prediction of news articles and fine-grained offensive analysis of social media comments. Results show that our methods and resources have proven to be fundamental for countering harmful online communication in Brazil.

11:00 – 12:30  
**Input session 4**

**Approaches for understanding harmful online content: traditional, computational, mixed methods**  
Moderation: Indira Sen, University of Konstanz

**Computational Approaches to Identifying and Mitigating Harmful Content Online**  
Libby Hemphill, University of Michigan

Abstract: Computational approaches, especially automated detection and mitigation, hold promise for improving online conversations. In this talk, I will review experiments using bystander bots and automated appeal mechanisms to de-escalate conflict online and machine learning approaches to detecting harmful content. Together, these projects illustrate how human-AI collaboration can improve detection models, reduce content moderator workloads, and reduce the overall incidence of harmful content.
Diana Rieger, Ludwig-Maximilians-University München
Tba

Anne Lauscher, University of Hamburg
Tba

Björn Ross, University of Edinburgh
Tba

14:00-15:30  **Work session**  
Factors influencing the quality of research on harmful online communication – towards a position paper / workshop summary paper  
Moderation: Katrin Weller, GESIS & Indira Sen, University of Konstanz

15:30-16:00  **Closing remarks**  
Katrin Weller, GESIS - Leibniz Institute for the Social Sciences
Appendix: Mini Poster Session - Lightning Talks
16.11.2023 17:00-18:00

Eduardo Barbabela, Andressa Liegi Vieira Costa
Digital Transformation and Political Communication: Exploring Negative Campaigning in Contemporary Democracies through Social Media Analysis

Thales Bertaglia, Rishabh Kaushal, Adriana Iamnitchi
The Monetisation of Toxicity: Analysing YouTube Content Creators and Controversy-Driven Engagement

Olga Bogolyubova
Dark Personality Traits and Deception in the Context of Online Dating

Janis Goldzycher
Data-Efficient Hate Speech Detection Via Natural Language Inference

Samuel Groesch
Infodemic on display: The role of visual misinformation during the Covid-19 pandemic. Exploring a visual topic modeling approach on German Covid-19 protesters and far rights on Telegram

David Grüning, Julia Kamin, Folco Panizza, Matthew Katsaros, Philipp Lorenz-Spreen
Designing for Prosociality: A framework and review of digital interventions for promoting online prosocial behavior

Ofra Klein
The Rhetorical Strategies of Radical Right Party Leaders in Legitimizing Hate Speech Online

Seda Mohul
The Online Representations of Immigrant Criminality on Twitter in Turkey

Lisa Oswald
Effects of Preemptive Empathy Interventions on Reply Toxicity among Highly Active Social Media Users

Pia Pachinger
A Multilingual Dataset for Target- and Span-Based Offensive Language Identification

Mugdha Pandya, Diana Maynard
Research Practices Influence Bias Too

Eirliani Abdul Rahman
The Anatomy of (In)Direct Swarming and Its Potential Threat to Democracies

Heidi Schulze, Simon Greipl, Julian Hohner, Patrick Schwabl & Diana Rieger
A Little Less Hate, But a Lot More Harm – Fear Speech as Strategic Borderline Communication

Alexander Sobieska
Towards a Differential Model of Online Radicalization - Insights from a Longitudinal, User-Focused Analysis of Radicalization Signals in Language on r/Incels
Tobia Spampatti, Ulf J. J. Hahnel, Evelina Trutnevyte, and Tobias Brosch
Psychological inoculation strategies to fight climate disinformation across 12 countries

Lea Stahel
Beyond sexism and racism: Unveiling the hidden discrimination of classist digital hate speech

Elisabeth Steffen
Visual Topic Modelling of Conspiracy Narratives

Peter Trolle, Luca Rossi, Christian Hardmeier
SafeNet: Challenges and Strategies in Combatting Online Hate Speech

Matti Wiegmann, Magdalena Wolska, Benno Stein, Martin Potthast
Introducing Computational Research on Trigger Warnings