	Wednesday – 02.04.	Thursday – 03.04.	Friday – 04.04.
9:00	ARRIVAL/REGISTRATION	Coffee	Coffee
9:30	Welcome	**Keynote 3 + QA: Maria Pawelec "Deepfakes for political manipulation and disinformation"**	**Keynote 5 + QA: Helen Fischer "Climate Change and Misinformation with societal consequences"**
10:00	**Keynote 1 + QA: Stephan Lewandowsky "Honest liars and the threat to democracy"**		
10:30		*BREAK*	*BREAK*
11:00	*BREAK*	Focus Session 3: Truth, Manipulation, and the Ethics of Misinformation	Session 5: Round Table
11:30	Focus Session 1: Politics, Bias, and the Influence of Misinformation		
12:00			
12:30			
13:00		LUNCH	LUNCH
13:30	LUNCH		
14:00		**Keynote 4 + QA: Segun Aroyehun**	Wrap-Up
14:30	**Keynote 2 + QA: Nadia Said "Understanding Sharing Behavior on Social Media Platforms"**		
15:00		*BREAK*	
15:30	*BREAK*	Focus Session 4: Detection, Influence, and the Spread of Misinformation	
16:00	Focus Session 2: Cognition, Trust, and the Psychology of Misinformation		
16:30			
17:00			
17:30			
18:00	**Dinner & Networking @ IWM**		

Keynote 1 + QA: Honest liars and the threat to democracy

Stephan Lewandowsky (University of Bristol)

Stephan Lewandowsky is a cognitive scientist interested in the pressure points between the architecture of online information technologies and human cognition, and the consequences for democracy that arise from those pressure points. This has led him to examine the persistence of misinformation and the spread of "fake news" in society, including conspiracy theories, and how platform algorithms may contribute to the prevalence of misinformation. He is also interested in the variables that determine whether or not people accept scientific evidence, for example surrounding vaccinations or climate science.

His research is currently funded by several agencies, such as the European Research Council, the EU's Horizon 2020 programme, and the UK Research Agency (UKRI, through the Centre of Excellence REPHRAIN). Because his research speaks to important contemporary issues, he works with policymakers, mainly at the European level, to make democracy more resilient to toxicity online. He also contributes to public debate through opinion pieces in the media and public engagements. In 2022, and again in 2023, he was identified as a highly cited researcher by Clarivate.

Keynote 2 + QA: Understanding Sharing Behavior on Social Media Platforms

Nadia Said (University of Tübingen)

Nadia Said is an interdisciplinary researcher specializing in the application of mathematical methods to understand psychological processes, with a particular focus on the (meta)cognitive underpinnings of misinformation and belief polarization. Her work explores how individuals process information, assess risks, and form opinions in the context of complex social issues such as climate change and artificial intelligence.

Her recent projects address topics such as metacognition in information processing, AI-related risk perception, and the psychological mechanisms behind susceptibility to misleading and manipulative content. Nadia's work has been supported by various grants and funding, including the Post-Doc Grant on Politicized Science and Media (2021), to develop a social media platform for studying individual sharing behavior on social media.

Keynote 3 + QA: Deepfakes for political manipulation and disinformation

Maria Pawelec (University of Tübingen)

Maria Pawelec studied Politics and Public Administration and Contemporary European Studies in Konstanz, Istanbul, Bath and Berlin. She then worked in the department for "International

Relations Europe and its Neighbors" of the Robert Bosch Foundation in Stuttgart. From 2016-2019, she worked as a Research Associate at the IZEW on the project "Ethical implications of IT-export to sub-Saharan Africa (ELISA)". Additionally, she supported the project Privacy-Arena in and coordinated the preparation of an expert report on behalf of the German Bundestag on the social and psychological impact of technological surveillance. In 2020, she worked on the project "Technological Innovation: Social Science and Ethical Analyses on Governance (TANGO)".
Since 2021, Maria Pawelec has been studying the "Ethical and societal implications of 'deepfakes' and opportunities for their regulation" within the project "Digitalization in dialogue (digilog@bw)". Since 2022, she is researching the prevention of digital desinformation campaign within the project PREVENT. In 2024, she also advised and supported the Federal Agency for Civic Education in creating the dossier "When perceptions deceive - Deepfakes and political reality".

Keynote 4 + QA: Conceptions of Truth: Computational Analyses and Societal Implications

Segun Aroyehun (University of Konstanz)

Segun Taofeek Aroyehun is a researcher specializing in natural language processing (NLP) and deep learning. His focus is on improving the understanding of natural language and addressing critical challenges in online communication. This includes efforts to combat hate speech, toxicity and extremism online, as well as advancing techniques for analyzing multilingual and code-switching texts.
In recognition of his significant research, Segun was awarded the prestigious Microsoft Research Latin America PhD Award in 2020. His research addresses pressing societal challenges by developing robust methods to prevent, detect and combat offensive content on social media platforms.

Keynote 5 + QA: Climate Change and Misinformation with societal consequences

Helen Fischer (University of Tübingen)

Helen Fischer investigates the role of metacognition, our insight into the reliability and the limits of one's own knowledge for beliefs about politicized science such as climate change or COVID-19. Her work illuminates the importance of metacognition for recognizing one's own errors in reasoning, such as motivated information processing and biased information transfer in social networks. In the winter semester 2023/24, she held a visiting professorship for Science and Society at the Karlsruhe Institute of Technology (KIT).
Helen Fischer earned her Ph.D. in cognitive psychology from the University of Heidelberg in 2016. After postdoctoral positions at the same university focusing on public perception of climate change, she received a DFG postdoc fellowship in 2019 to work at the Stockholm Resilience Center

in 2019-2020. She was a visiting researcher at the Max Planck Institute for Human Development from 2020 to 2022. Since 2022, she has been conducting research at the Leibniz Institute for Knowledge Media, Tübingen. In 2023 she vistited the University of Waikato, New Zealand, to investigate the longitudinal connection between social media use and climate change beliefs in a Fritz Thyssen Foundation-funded project.

Focus Session 1: Politics, Bias, and the Influence of Misinformation

Talk 1: Christoph Abels (University of Potsdam)

Driving democratic backsliding: Misinformation and its role in democratic decline

Democracy is in decline worldwide, with 71% of the world's population now living in an autocratic state. This trend is not confined to younger democracies—it is unfolding in established liberal democracies, including the United States. In these systems, democratic erosion is often instigated and sustained by political elites—elected officials, party leaders, bureaucrats—who push the boundaries of their power. Left unchecked, this process can reach a tipping point, after which democracy gives way to autocracy. However, elites do not operate in isolation. An attentive public can constrain them, either by withholding support or voting them out of office. But when the public is misled, indifferent, or actively supportive, norm violations escalate, and the guardrails of democracy weaken.

Misinformation is a critical enabler of this process. Political elites can strategically deploy misinformation to cultivate a compliant or complicit public—one that either tolerates norm violations or demands them. This dynamic is unfolding in real time in the United States, where misinformation has been used to justify the dismantling of key institutions. It also played a central role in undermining the peaceful transfer of power in 2021, culminating in the January 6th insurrection.

In this talk, I will examine the role of misinformation in democratic backsliding, using the drift-to-danger model to illustrate how misinformation erodes democratic safeguards and accelerates institutional decline. I will also outline behavioral science-informed countermeasures designed to strengthen democratic resilience in the face of strategic manipulation.

Misinformation democratic backsliding norm violations Narratives Neuroimaging

Talk 2: Gafari Lukumon (African Business School, University of Mohammed VI, Polytechnic, Rabat, Morocco)

Adversarial collaboration among fact-checkers: bipartisanship outweighs partisanship in preferences and trust towards news

Have politics and news become so polarized that people prefer co-partisan fact-checkers over bi-partisan fact-checkers? Or despite a polarized political climate, do people nevertheless still prefer bipartisan fact-checking? In this study, we investigated individuals' attitudes toward ideologically heterogeneous (i.e. politically mixed) vs. homogenous fact-checking groups. Across two pre-registered studies (N = 876), we investigated preferences (and motivations) for consulting such groups as well as trust in them. Left- and Right- leaning American participants exhibited a strong preference for bipartisan fact-checking groups, with the greatest preference for political diversity

among left-leaning participants. Notably, both left leaning and right leaning participants trusted heterogeneous fact-checking groups as much as their own but expressed lower trust in the opposing political party. These studies suggest that even in a highly polarized political climate, people prefer a diversity of viewpoints to be included in the fact-checking process. This insight may offer a new perspective on overcoming partisan divides in information consumption.

Fact-checking adversarial collaboration bipartisan news trust

Talk 3: **Juan Vidal-Perez** (UCL)

Biased Misinformation Fosters Biased Beliefs

Systematically biased information, such as politically slanted news or selective media coverage, has become a prominent source of misinformation, fueling political polarization, the persistence of false beliefs, and the spread of conspiracy theories. Can we maintain accurate representations of our environment when we are bombarded with biased information? Here, we use a reinforcement learning paradigm to investigate how people process biased information. We developed a multi-armed bandit task where participants (n=200) received feedback about reward-outcomes from potentially biased sources: some favorable (overestimating rewards), some unfavorable (underestimating rewards), and some unbiased (providing accurate reports). Participants were first given the opportunity to learn about the bias of different sources through a comparison of their feedback to "ground truth" choice-outcomes. Subsequently, access to true choice outcome was withheld so participants had to use this learned source-bias knowledge to correct feedback from those sources to form accurate beliefs about the value of bandits. Our results reveal two key findings: First, individuals correct biased information in the right direction, by inflating unfavorable feedback and dampening favorable feedback. However, they only correct feedback partially (i.e., they under-debias), leading to the propagation of source biases into beliefs biased in the same direction. Second, participants frequently misclassified unbiased sources, often perceiving them as biased in the opposite direction of concurrently presented biased sources (e.g., misjudging an unbiased source as unfavorable when concurrent sources are favorable). These intertwined effects of under-debiasing and misclassification illustrate how biased information can shape beliefs and contribute to the spread of misinformation.

Bias Reinforcement learning Human Behavior Computational Modeling

Talk 4: **Gautam Kishore Shahi** (University of Duisburg-Essen, Germany)

Too Little, Too Late: Moderation of Misinformation around the Russo-Ukrainian Conflict

We examine the role of Twitter as a first line of defense against misinformation by tracking the public engagement with, and the platform's response to, 500 tweets concerning the Russo-Ukrainian conflict which were identified as misinformation. Using a real-time sample of 543475 of their retweets, we find that users who geolocate themselves in the U.S. both produce and consume the largest portion of misinformation, however accounts claiming to be in Ukraine are the second

largest source. At the time of writing, 84% of these tweets were still available on the platform, especially those having an anti-Russia narrative. For those that did receive some sanctions, the retweeting rate has already stabilized, pointing to ineffectiveness of the measures to stem their spread. These findings point to the need for a change in the existing anti-misinformation system ecosystem. We propose several design and research guidelines for its possible improvement.

Misinformation Russia-Ukraine conflict Narrative Content Moderation Twitter

Talk 5: **Raphael Zähringer** (University of Tübingen, English Department)

The Contingencies of Review Culture in Social Media

In an era of increasing social complexity, customer ratings and reviews have become vital mechanisms for navigating an overwhelming world. Review culture on social media fosters a sense of security by transforming subjective experiences into quantifiable metrics, allowing consumers to make seemingly informed decisions. However, these very mechanisms introduce new layers of uncertainty, as they are subject to manipulation, emotional bias, and systemic distortions such as review bombing. This paper explores these contingencies through the case of The Shed at Dulwich, a fictitious restaurant that climbed to the top of TripAdvisor's rankings despite never serving a single real customer. The viral hoax exemplifies how social media amplifies trust in aggregated opinions while simultaneously undermining their reliability. Framed through the lens of systems theory, particularly the concepts of contingency and social complexity, this paper argues that review culture operates as a paradox: it offers stability in decision-making while proliferating new uncertainties. Rather than providing objective assessments, online ratings construct a reality based on iterative social feedback, where credibility is performative rather than substantive. The Shed at Dulwich incident highlights the extent to which digital reputations are contingent on belief rather than truth, raising critical questions about the epistemological foundations of trust in online environments. By analyzing review culture as a system that both simplifies and destabilizes reality, this paper illuminates the paradoxical role of digital platforms in shaping contemporary social knowledge.

Contingency complexity reviews TripAdvisor

Talk 6: **Theresa Hermann** (University of Tübingen)

Deconstructing explanatory videos in history education

Explanatory videos have become increasingly popular in educational contexts (Medienpädagogischer Forschungsverbund Südwest, 2020). However, their use in history education is controversial, as these videos often present simplified or biased narratives (Anke, 2017; Arnold & Barth, 2024). The videos often follow a positivist pattern which omits scientific controversies. In some cases, this leads to presenting a false narrative thus raising concerns about misinformation (Zenthöfer, 2024). Despite their widespread consumption, research on their role in historical learning remains limited and no standardized framework exists for their systematic analysis. My PhD project, which is still in its early stages, aims to synthesize existing research from

history education to develop a structured framework for analysing history themed explanatory videos. The framework, a first draft of which will be presented at the workshop, will be implemented in an intervention to help students deconstruct video content and identify potential biases, omissions, and misleading claims. Drawing on the principles of lateral reading (Wineburg & McGrew, 2017) and civic online reasoning (McGrew et al., 2018), this approach will be implemented in a lesson plan to investigate, whether it can foster digital literacy and encourage students to engage more critically with historical narratives.

Anke, J. (2017, Juni 29). Wissen2go – History on YouTube. PublicHistoryWeekly. https://public-history-weekly.degruyter.com/5-2017-25/wissen2go-teacher-centered-instruction-on-youtube/

Arnold, K., & Barth, S. (2024). Erklärvideos im Geschichtsunterricht. Geschichte lernen, 217. https://www.friedrich-verlag.de/mein-friedrich/#/article/17221

Medienpädagogischer Forschungsverbund Südwest. (2020). JIM-Studie 2020. Jugend, Information und Medien. Basisuntersuchung zum Medienumgang 12- bis 19-Jähriger. https://www.mpfs.de/studien/jim-studie/2020/

Zenthöfer, J. (2024, April 29). Bei „MrWissen2Go“ werden englische Stallburschen zu Hitlerjungen. FAZ.NET. https://www.faz.net/aktuell/feuilleton/medien/bei-mrwissen2go-werden-englische-stallburschen-zu-hitlerjungen-19687024.html

Explanatory Videos | History Education | Deconstruction | Civic Online Reasoning | Narrative Bias

Talk 7: **Simone Zanello** (University of Tübingen)

A Post-Structuralist Account of Truth and Information.

In a regime of post-truth (Kalpokas, 2019), the noble task of fighting disinformation could give the impression that not only is still possible to distinguish between True and False but also that the possibility of a faithful representation of Truth still exists or has ever existed. The aim of my presentation will be instead to challenge the couple Information/Disinformation as a dichotomy of Truth/Falsehood.

Using Michel Foucault's analysis of parrhesia, I’ll begin by showing how saying the truth has historically been comprehended as a multifaceted act, that in the Western context functioned as an obligation and/or a commitment to a personal relationship with the truth (Foucault, 2016, pp. 21–76). Subsequently, I’ll move to the concept of Maximal Ideal Information formulated by Gilles Deleuze and Felix Guattari, emphasizing the role of redundancy as a device of noise-reduction for the transmission of mots d’ordre, in turn functioning as markers of power (Deleuze & Guattari, 2016, pp. 95–96). Finally, recalling the process of precession of the simulacra (Baudrillard, 2024), I’ll conclude by pointing to the problematic relationship of Truth and information through the not-having-taken-place of the Gulf War (Baudrillard, 2002).

Bibliography

Baudrillard, J. (2002). The Gulf War did not take place. Indiana Univ. Press.

Baudrillard, J. (2024). Simulacres et simulation. Gallimard.

Deleuze, G., & Guattari, F. (2016). Mille plateaux. Les Éditions de Minuit.

Foucault, M. (2016). Discours et vérité—Précédé de La parrêsia. Librairie Philosophique J. Vrin.

Kalpokas, I. (2019). A Political Theory of Post-Truth. Springer International Publishing.

Truth Post-Structuralism Simulacra Philosophy of Information Political Theory

Talk 8

Focus Session 2: Cognition, Trust, and the Psychology of Misinformation

Talk 1: **Delaram Sadeghzadeh** (University College London)

Misinformation as a Memory Control Process: Distinguishing Real from Imagined

Misinformation presents a significant cognitive challenge, requiring individuals to distinguish between true and false information while drawing on prior knowledge. Rooted in Bartlett's (1932) theory of reconstructive memory, this study examines how memory control processes influence information evaluation. Our previous exploratory studies suggested that familiarity with news and reality monitoring (RM) ability—the capacity to differentiate perceived from imagined events (Johnson & Raye, 1981)—contributes to detecting fake news. However, as these studies relied on real-world news content, prior exposure was uncontrolled, making it difficult to isolate the memory mechanisms underlying misinformation processing. The current study systematically manipulates prior knowledge and familiarity to refine our understanding of memory control processes in information evaluation.

Adapting the Deese-Roediger-McDermott (DRM) paradigm (Roediger & McDermott, 1995), participants read short fictional narratives embedded with semantically related word lists designed to induce false memories. In the retrieval phase, they evaluate statements based on:

Veracity Judgment – Determining whether statements are true or false.

Gist-Based RM – Differentiating between old (previously mentioned) and new (novel but plausible) information.

Verbatim RM – Identifying lures (critical, unpresented words) vs. non-lures (explicitly presented words).

We also collect confidence ratings, demographics, and personality measures to explore individual cognitive and metacognitive differences. Future extensions will incorporate functional near-infrared spectroscopy (fNIRS) neuroimaging to examine the neural underpinnings of reality monitoring and misinformation evaluation.

Misinformation Susceptibility Reality Monitoring False memory Neural Basis of Memory Control Metacognitive Sensitivity

Talk 2: **Gabriel Braun** (Tel Aviv University)

Shared Disbelief, Shared Belief: Disbelief Context and Actual Belief as Drivers of Interpersonal Neural Synchronization

Despite living in an era where the mere concept of truth is increasingly contested, the cognitive processes underlying the processing of information we believe or disbelieve in remain largely unexplored. In this fMRI study, we examined narrative processing through two factors of belief: Belief context- contextual information that provides initial indications of truthfulness, and actual belief- the truth value ultimately assigned to the narrative. Participants (N=48) listened to two

narratives where context either supported or discredited the speaker's account of the events. After each narrative, they were asked to report their actual belief in the narrative. To investigate the effects of (dis)belief on narrative processing, we analyzed neural synchronization using inter-subject-correlation analysis and inter-subject representational similarity analysis. We successfully decoded (dis)belief context by modeling neural synchronization patterns, despite an actual "belief-bias" at the behavioural level. This indicates a unique neural pattern related to each belief context. Notably, a dissociation emerged between belief context and actual belief both in localization and in opposite synchronization patterns: Disbelief context led to higher synchronization compared to belief context, in areas involved in conflict control, while actual belief led to higher synchronization compared to actual disbelief, in areas associated with the mentalizing network. These results highlight the dissociable aspects of (dis)belief, which, while interconnected, correspond to distinct cognitive processes. Specifically, our findings suggest that disbelief context activates perceptual mechanisms related to conflict monitoring and resolution that underly interpretation of events. In contrast, actual belief promotes a more coherent interpretation, resulting in higher synchrony among participants.

Belief Distrust fMRI Narratives Neuroimaging

Talk 3: **Margarita Pavlova** (New Bulgarian University**)**

Drivers of Misinformation: **What Can We Learn from Analogy and Memory Research?**

In an era of information abundance, one is required to adequately screen, interpret, and critique sources of information. Analogical reasoning – the ability to notice patterns of similarity across disparate domains – interplays with memory such that people can blend information from multiple sources. Classical models of human memory predict that people blend elements of events that share surface similarity, such as perceptual or semantic commonalities (McClelland, 1995; Metcalfe, 1990; Murdock, 1995). However, a vast body of research shows that people can and sometimes even prefer to blend information from events that share relational commonalities (Feldman & Kokinov, 2009; Kokinov & Petrov, 2001; Pavlova & Kokinov, 2014; Raynal et al., 2020; Trench & Minervino, 2015). Thus, analogical reasoning can influence encoding, storing, and retrieving of information and can potentially guide one's attention to past examples that can mitigate or corroborate misinformation. Therefore, having a better understanding of the mechanisms, benefits, and limitations of analogical reasoning, can be fruitful in efforts to mitigate misinformation. In this presentation, I will summarize the current research on the interplay between memory and analogy and discuss theoretical implications and practical considerations of how analogical reasoning relates to the psychology of misinformation.

Analogy memory relational retrieval

Talk 4: **Isabella Orpen** (Cardiff University)

Consuming Conspiracies: Understanding heterogeneity of consumers of conspiracy theories

The coronavirus pandemic and the rise of populism globally have exemplified the power of conspiracy theories and their detrimental impacts on public-health, democracy and security. The vast and ever evolving information landscape means humans are required to make more and more complex decisions regarding their consumption of information.

This research examines conspiracy mentality in the UK population to create a topology of the pathways toward and away from high conspiracy mentality. Although recent research has made great strides into identifying conditions which impact vulnerability and resilience to conspiracy theories further investigation is needed to understand how these conditions combine and impact one another to impact the manifestation of conspiracy mentality in society.

To address these research questions this thesis employs and combines the motivational model and social identity theory to understand how different conditions might be interconnected. It utilises a survey of 500 respondents in the UK to establish the extent of conspiracy mentality in society and using fuzzy set qualitative comparative analysis examines how the configurations of conditions combine to relate to conspiracy mentality. This highlights the heterogenous pathways toward and away from conspiracy mentality. From these heterogenous pathways a typology can be created which emphasises the diverse types of consumers of information in relation to conspiracy theories. By segmenting the audience for conspiracy theories in this way the different behaviours, attitudes and beliefs can be more acutely assessed.

Conspiracy theory FsQCA consumer psychology social identity motivational model

Talk 5: Eva Rudholzer (Leibniz-Institut für Wissensmedien | Everyday Media Lab)

Skimming Content on Social Media: Are Users Susceptible to Misinformation?

Social media has become a primary source of (scientific) information, but its design encourages rapid scrolling and skimming, making it more difficult for users to identify the source and verify the credibility of the content. Prior research suggests that source cues (sender characteristics) influence credibility evaluations, often more than content type. However, most studies have assessed the credibility of each post one after the other rather than in the context of a feed. Therefore, it remains unclear whether users rely more on source cues or content type when evaluating credibility while scrolling.

To investigate this, we will conduct a preregistered online experiment where participants scroll through a feed with posts from different senders (role: scientist vs layperson vs journalist, within) sharing various types of science-related content (content-type: peer-reviewed paper vs preprint vs misinformation, within). Since users may not know who said what after scrolling through a feed, we examine whether they confuse the roles of senders and hypothesise that source cues have a more negligible influence. This might have implications for users' susceptibility to misinformation: If source cues are not dominant anymore, the question arises as to whether users can still distinguish credible from non-credible content.

Our findings will clarify how social media affects users' credibility judgments for scientific content. Misattributing credibility could lead to misplaced trust in unreliable sources and scepticism toward legitimate scientific findings, which could ultimately undermine public trust in science. Understanding these mechanisms is crucial for developing interventions that enhance users' critical evaluation of information.

science communication social media perceived credibility misinformation

Talk 6: **Laura Burkhardt** (University of Tübingen)

Understanding sharing behavior on social media platforms: **How the ability to discern between manipulative and non-manipulative content influences sharing manipulative and false information**

Social media has shaped our world in the last decades. The initial enthusiasm, however, transformed into concerns about the dangers of unchecked sharing of false information. This has prompted a large body of research about the spread and impact of false information. Recent research showed that misleading information (implied misinformation) rather than false information is of much more concern. To gain a deeper understanding of what predicts sharing of misleading information we adapted the Manipulative Online Content Recognition Inventory (MOCRI) for use in the German context. The adaptation involved translating and culturally aligning the MOCRI items and testing them on a German sample. The adapted MOCRI scale was evaluated for its reliability and utility in predicting engagement with different types of content, laying the groundwork for future behavioral investigations into misinformation sharing behaviour.

MOCRI social media manipulative information sharing behaviour

Talk 7: **Marija Neralić** (Institute of Social Sciences Ivo Pilar, Zagreb)

Developing and piloting climate change disinformation videos in the context of DISINFO climate project

The DISINFO climate project (Next Generation EU 2024-2027) is focused on the role of dispositions and individual characteristics (personality traits, cognitive styles, attitudes) and exposure to (dis)information on climate change (CC) attitudes and behaviors. The main aim is to experimentally investigate how the combination of individual characteristics and exposure to disinformation shapes CC attitudes and behaviors. The first study (N1 ≈ 4000) will examine the effect of exposure to CC disinformation on attitudes and behaviors, considering different participant profiles. The second study (N2 ≈ 3250) will examine the effects of techniques to mitigate the impact of disinformation while considering different participant profiles. Besides scientific contributions, the project will have practical implications for public policymakers in combating disinformation.

The presentation will focus on preparatory work for the first study, which included several challenges related to selecting, developing, and piloting disinformation for experimental manipulations. I will present the most common categories of CC disinformation posts and articles in Croatia and examples of developed and preliminary piloted experimental manipulations (short videos in the form of social media posts created with the help of AI).

climate change disinformation social media artificial intelligence

Talk 8: **João Pedro Silva Vieira** (University of Porto)

PhD Project: **Uncovering the cognitive mechanisms of inoculation against misinformation**.

The spread of online misinformation (aka fake news) is growing and poses significant challenges to science, health and democracy. Among strategies to reduce misinformation, pre-emptive exposure to misinformation examples (aka inoculation) is more effective than post-exposure correction (debunking). Recently, gamified versions of an inoculation intervention teaching common misinformation techniques have been successfully implemented and tested. However, the cognitive mechanisms behind inoculation effects are still ill-specified. To fill this gap, we will examine how a well-known gamified intervention (the Bad News game) protects against misinformation across three work packages (WP). WP1 will contrast two mechanisms of change: improved fake news detection vs. increased skepticism. WP2 will evaluate how inoculation changes the allocation of attention to fake news cues using eye-tracking. Finally, WP3 will uncover how manipulations aimed to increase learning (gamification and memory reminders) affect the longevity of inoculation. This project will gather crucial knowledge to design better immunization tools against misinformation.

Inoculation Fake News Eye-Tracker Scepticism

Focus Session 3: Truth, Manipulation, and the Ethics of Misinformation

Talk 1: **Anne Hausknecht** (Swansea University)

Preserving Trust in User-generated Evidence in an Era of Deepfakes

The rapid development of deepfake technology (meaning machine learning technology to create hyper-realistic images, videos, or audio recordings) has caused fears of the emergence of a new post-truth-era. Deepfakes could pose a particular threat to efforts of documenting human rights violations and preserving evidence for legal accountability processes, which increasingly rely on user-generated evidence ("UGE", i.e. information captured by ordinary users on their personal devices and used in legal adjudication). Scholars have suggested that deepfakes might affect efforts of documenting mass atrocities in three ways: authentication of evidence becomes more difficult; real evidence is decried as fake (called the liar's dividend); and by causing general distrust in all UGE (the impostor bias). To date, no study has tested to what extent deepfakes actually impact efforts of investigating and prosecuting mass atrocities.

Examining international crimes cases in Finland, Germany, the Netherlands, and Sweden, my presentation fills this gap by investigating which steps practitioners and investigators have and should take to address the challenges posed by deepfakes. I explore how the existence of deepfakes has changed the process of authenticating UGE. I then test if the liar's dividend has been used as a strategy to undermine trust in evidence. Based on a series of interviews with legal practitioners and open-source investigators, I further examine whether the existence of deepfakes has or is likely to lead to general distrust in all evidence.

Trust deepfakes user-generated evidence liar's dividend

Talk 2: **Fang Zhao** (Ostbayerische Technische Hochschule Regensburg)

Misinformation Belief Correction: Before, During and After Contact to Misinformation

Belief in misinformation can lead to poor reasoning. Even after correction, its influence can still linger and can significantly impact people's decision-making. This literature review covers psychological works on definitions, types, related terms of misinformation. It explains the drivers of misinformation beliefs and its continued influence after correction. It compares psychological theories and paradigms on how people correct misinformation belief before, during and after contact to misinformation. Effective strategies are discussed to counter misinformation, such as inoculation techniques for prebunking, fact-checks during contact, and controlled information retrieval for debunking. Finally, the review discusses limitations of theories and methods and provides implications for laypeople, educators, policy makers, and information consumers.

inoculation theory elaboration likelihood automatic information retrieval controlled information retrieval

Talk 3: **Gerrit Anders** (Leibniz Institut für Wissensmedien)

Can information theory help us understand the spread of misinformation?

In the digital landscape, misinformation spreads rapidly, posing significant challenges to public discourse and decision-making. This work explores how a novel framework based on information theory can be developed to better understand and analyze the propagation of misinformation. We will explore the foundational concepts and potential pitfalls of applying information theory to this issue.

We argue that misinformation is not merely noise interfering with communication channels but information with altered truth values and content. This perspective necessitates extending classical information theory to incorporate semantic content and the concept of truth. By considering misinformation as information rather than noise, we explore approaches to account for informational content and its distortion.

When discussing the pitfalls of adapting information theory to encompass truth values, we highlight challenges in quantifying semantic content and addressing the non-binary nature of truth. Additionally, we explore how misleading information and narratives can manipulate and undermine truth, even when they cannot be strictly evaluated by truth values. In order to do so, we highlight shared information as a crucial concept in investigating information distortion and its propagation through networks.

By building on the mathematical rigor of information theory, our framework aims to provide a robust analytical tool for understanding misinformation. By quantifying key aspects of misinformation propagation and distortion, this approach holds promise for enhancing our understanding and developing effective mitigation strategies regarding misinformation.

information theory misinformation mathematics

Talk 4: **Stef Hankel** (Radboud University | Centre for Language Studies)

The hand behind misinformation: A systematic review on the linguistic, structural and rhetorical characteristics of online misinformation

As misinformation continues to spread online, overwhelming the public with false and misleading narratives, it is becoming increasingly difficult to discern fact from fiction. Consequently, scholars are developing interventions that try to help people recognize misinformation by its linguistic characteristics. To achieve this, researchers rely on both computational as well as on qualitative analyses to identify 'misinformation' features. Although these methods are complementary, the former generally reveals hidden patterns that are difficult to spot for human readers, whereas the latter typically relies on uncovering idiosyncratic features, for example in rhetoric. This systematic review examines peer-reviewed studies to address the following research question: Which linguistic, structural and rhetorical features characterize misinformation? We provide an in-depth overview on all these identified textual characteristics of misinformation, focusing particularly on how these are related to misinformation type (e.g. conspiracies or fake news), to online platforms

(e.g. Facebook, X or news websites) and to the specific format in which they appear (e.g. news articles or social media comments). Additionally, we assess which of these characteristics are potentially perceptible through human observation. Current gaps and avenues for future research as well as the implications of these findings on how to improve literacy skills are discussed.

Misinformation disinformation linguistics rhetoric online platforms

Talk 5: **Asya Achimova** (University of Tübingen)

Persuasion without accountability: **the case of projection inferences**

If a speaker wants to communicate new contents to the audience, she can do so via an assertion - a type of speech acts associated with declarative sentences, such as “The government tried to form a coalition”. Speakers are normally held accountable for the contents they assert. Thus, a listener skeptical about the willingness of the government to form coalitions could deny this content by saying “That is not true”. However, language offers subtle ways to embed contents in an utterance without the speaker being accountable for it. For example, the speaker could say “The government failed to form a coalition”. For the utterance to make sense, the audience has to accept that the government at least tried to do so. While in the case of assertions, the listener may challenge the contents of the assertion, it is not possible to challenge the presupposed content in the same way. In a psycholinguistic experiment (n = 200), we asked participants to read stories that presented historic made-up facts. We manipulated the way we presented the information (via assertions vs. presuppositions), as well as the types of presupposition triggers. We demonstrate that for some types of presupposition triggers (factive verbs and counterfactual conditionals), participants find the presupposed contents as plausible as that of assertions. Thus, presuppositions ofter speakers an opportunity to make the audience infer contents without the speaker being accountable for that contents. We discuss these findings in the cognitive framework of predictive coding.

Language presupposition inference accountability manipulation

Talk 6: **Jörn Wiengarn** (TU Darmstadt)

Hidden value judgments in disinformation detection technologies

Most of the technical literature on disinformation detection treats the challenge of reliably identifying disinformation – while avoiding the misclassification of legitimate information –as a purely epistemological or technical problem. In this talk, I challenge that framing. I argue that, rather than being neutral, the process of identifying disinformation is inherently value-laden. Implicit value judgments shape the classification of information as disinformation and subsequently the technologies designed to detect it.

I will develop this argument in three steps. First, I will provide a systematic overview of the dimensions in which distinguishing between disinformation and accurate information becomes ambiguous. Second, I will reframe this ambiguity as an ethical issue rather than a purely epistemological one. The question is not just what de facto counts as disinformation, but also what we consider legitimate for public discourse. The act of labeling information as disinformation

implicitly reflects normative priorities about the public sphere. Finally, I will show how these normative assumptions are embedded in the design choices of disinformation detection technologies. To illustrate this, I will draw on examples from the literature on fake news detection, demonstrating how technical systems enshrine underlying value judgments.
By making these implicit values explicit, my talk will highlight the ethical stakes of disinformation detection and invite a critical reassessment of how we design and deploy these technologies.
Bias Reinforcement learning Human Behavior Computational Modeling

Talk 7: **Ruvindra Sathsarani** (University of Tübingen)

Truth, **Lies**, **and Algorithms**: **Pretended Personalities and the Ethics of Digital Misinformation**.

Misinformation on social media and its ethical dilemmas have become a crucial topic in recent years. Lauren Oyler's 2021 novel Fake Accounts offers a sharp critique of this phenomenon, exploring the blurred boundaries between truth and deception. When the novel's protagonist discovers her boyfriend's secret online identity—where he propagates various conspiracy theories—she is confronted with the instability of digital identity and the ease with which misinformation spreads. Oyler portrays social media as a space where personal identity is fragmented, and truth is easily manipulated, leading to socio-political turmoil.This study argues that Fake Accounts not only critiques the spread of false information but also examines the motivations behind it. Drawing on Baudrillard's concept of simulacra, which explores how reality is distorted and dismissed through semantic signs, this analysis situates the novel within a broader culture where social media erodes authenticity. By interrogating issues of personal agency, freedom of expression, and the moral ambiguity of digital discourse, this paper examines how misinformation operates beyond individual deception, shaping public perception and ethical responsibility. Ultimately, by positioning Fake Accounts within contemporary debates on digital media, this study questions the role of ethical conduct in a virtual landscape where reality is constantly rewritten, blurred, and manipulated for personal gain.
Social media discourse Misinformation Ethical conduct

Talk 8: **Michael Klenk** (TU Delft)

The ethics of (online) manipulation - Beyond good and after evil

As AI systems become increasingly integrated into information ecosystems, concerns about deception and manipulation grow. However, the challenge of misinformation is not merely the result of malicious intent. In my contribution, I will discuss a more nuanced and insufficiently explored risk: the subtle but pervasive manipulation embedded in AI's sycophantic tendencies—its inclination to flatter, affirm biases, adapt to the user's perspective, and reinforce existing beliefs. Sycophancy is evident in major AI applications such as ChatGPT and Claude, likely stemming from

reinforcement learning from human feedback. Yet, its ethical implications and impact on misinformation and democracy remain underexplored.

I will argue that sycophancy constitutes a form of ethically problematic manipulation, posing risks to epistemic processes essential for democracy, such as truth-seeking and informed decision-making. Drawing on philosophical literature on deception and manipulation, as well as technical research on AI deception, I will document the problem and argue for its significance. One of the upshots of the talk – which I hope will inspire discussion amongst the audience – is that deception, understood as intentional misleading and thus akin to disinformation, is in some ways less of a threat than manipulation, which is characterized as 'mere' indifference to truth, and of which sycophancy is a prime example.

By reconceptualizing manipulation beyond explicit deception, I aim to provide a fresh perspective on misinformation risks in AI-mediated environments. While solutions remain uncertain, my contribution will serve as a starting point for critical discussion on mitigating AI-driven epistemic distortions.

AI deception manipulation sycophancy misinformation democracy

Focus Session 4: Technology and Methods to Identify and Prevent Spread

Talk 1: Lulu Ito (Keio University)

Unmasking Digital Puppeteers: Revealing the Key Features for Social Bot Detection in the Japanese Context

Recent geopolitical instability is no longer just marked by territorial aggression but increasingly by cognitive warfare, where the lines between peacetime and conflict are blurred. The rise of social media platforms has provided malicious actors with the tools to exploit algorithmic vulnerabilities through automated systems, especially during political events. While computational propaganda has become a global concern, Japan has historically had limited exposure to such influence operations compared to other developed nations. However, 2024 proved to be a pivotal year, with three major elections heavily impacted by social media dynamics. This emerging threat to Japan's information ecosystem underscores the need for further research, which remains constrained by institutional limitations on war-related research funding and restrictive data collection policies from major social media platforms.

To address these challenges, we applied dimensional reduction techniques to condense a feature space of 13,767 attributes into 6 highly discriminative features while minimizing computational overhead. This approach resulted in outstanding classification performance, with an AUC-ROC of 0.999 and an F1-Score of 0.998. Furthermore, we conducted an in-depth analysis of the primary features contributing to social bot detection, establishing significant correlations between algorithmic indicators and their real-world behaviors. These findings offer valuable insights for advancing social bot detection and understanding the evolving landscape of computational propaganda in Japan.

Cognitive Warfare Computational Propaganda Social Bot Feature Selection

Talk 2: Huiyun Tang (University of Luxembourg)

The Potential of AI-Driven Prebunking in Combating Misinformation

The rapid spread of misinformation poses a threat to democracy and public trust. To address this issue, various digital interventions have been developed to slow its propagation. These interventions typically fall into two categories: debunking, which corrects falsehoods after exposure, and prebunking, which builds psychological resistance beforehand. While debunking is a necessary strategy, it may not be sufficient due to misinformation spreads faster and deeper than the truth. In addition, the continued influence effect of misinformation makes it challenging to mitigate its influence, making proactive prebunking approaches increasingly important.

Prebunking, based on inoculation theory, exposes individuals to weakened forms of misinformation, helping them develop cognitive immunity against future encounters. Prebunking is based on educational approach, include games. However, There are concerns around current

approach: misinformation tactics evolve rapidly, requiring constant updates; cultural and linguistic differences make it difficult to create universally effective interventions; and the impact of prebunking fades over time, highlighting the need for ongoing reinforcement.
Recent advancements in AI present new opportunities to address these challenges, enabling more interactive, replayable experiences. For instance, AI-driven Socratic questioning can prompt users to critically analyze their reasoning, fostering deeper cognitive engagement. Additionally, LLM-powered game environments can simulate dynamic misinformation scenarios, allowing players to experience realistic decision-making processes beyond predefined choices.
With the widely and successfully explored debunking efforts—such as credibility indicators and automated fact-checking, its potential for explore AI-driven prebunking solutions. This workshop can explore how AI can be leveraged to create adaptive, engaging, and scalable prebunking interventions.
Prebunking media literacy AI

Talk 3: **Nicole Antes** (Leibniz-Institut für Wissensmedien | University of Tübingen)

Can community notes serve as correction for misinformation

The Continued Influence Effect (CIE) - the persistent belief in misinformation despite corrections – is a challenge to interventions aimed at reducing the risk of harmful content, with social media playing a crucial role in spreading and correcting misinformation. This study explores the effectiveness of community notes, a crowdsourced fact-checking tool by X, in mitigating the CIE across neutral (e.g., food recall) and socially relevant content (e.g., gender quota). A German sample (N = 542) read narratives presented in tweets with one tweet containing misinformation. This tweet was later presented with community notes offering either a simple retraction or a correction with an alternative explanation. For neutral content, both retraction types reduced the CIE, with alternative explanations being more effective. For socially relevant content, only corrections with alternative explanations reduced misinformation use. These results suggest community notes can counter misinformation but may be moderated by ideological factors in socially relevant contexts
continued influence effect community notes social media correction

Talk 4: **Ruben Lamers James** (Swansea University)

Impact of Perceived Source on Belief Updating for Human Deepfake Detection

As individuals struggle to reliably detect deepfakes, with studies finding limited success in improving people's detection abilities, alternative strategies to reduce their susceptibility are needed. While AI detection models and crowd-based judgments show promise in improving detection accuracy, limited research has been undertaken on how people might use the information obtained through these routes to improve their deepfake detection. This study investigates whether the different sources of deepfake warnings affect advice uptake and

detection accuracy in a deepfake detection task. By flagging videos predicted to be deepfakes based on majority judgments from a prior study, we provide the first test of whether providing crowd-sourced veracity judgments enhance human deepfake detection. Participants, assigned to one of four groups first judge the authenticity of real and deepfake videos, and report their confidence. In the experimental phase, flagged videos (predicted to be deepfakes by the majority judgment) are accompanied by warnings, with their sources attributed to either an AI detection model, crowd consensus, or independent fact-checkers. A control group received no warnings. Building on misinformation research, we predict participants to be most receptive to the independent fact-checker warnings, followed by AI and then crowd-sourced judgments. Additionally, we will investigate whether providing these warnings lead to an illusory-truth effect, where non-labelled videos are more likely to be believed to be real. This study will also investigate the impact that participants' confidence and their Active Open-Minded Thinking tendency has on how likely they are to update their beliefs in response to receiving warnings.

Deepfakes Misinformation Wisdom of the Crowds Fact-checking AI

Talk 5: **Pauline Frick** (University of Tübingen, Leibniz-Institut für Wissensmedien)

Addressing Climate Change Misinformation: How Source Credibility and Graphs Influence Belief and Engagement with Climate Change Posts

Although research links human activity to climate change (Cook et al., 2016, IPCC, 2022), false narratives persist, largely shaped by mainstream media rather than scientific sources (e.g., Ecker et al., 2022). Effective science communication is crucial to counter false narratives and drive informed action. In two studies, we investigated how source credibility and the presence of graphs influence belief in climate change posts and liking and sharing intentions. As an exploratory variable, we included the general trust in science/scientists as moderator. We created 40 social media posts about climate change and varied whether these posts included a high or low credible source and whether they were presented with or without a graph. In Study 1 (N = 160), participants rated these posts on a 1-5 scale, indicating their belief in the information. In Study 2 (N = 167), participants rated the same posts on 1-5 scales, assessing their likelihood of liking and sharing. Regarding belief in information, high source credibility led to a higher belief in the information, while scientific graphs did not positively influence belief. Regarding liking and sharing intentions, the presence of graphs positively influenced both, but this effect was fully moderated by trust in science/scientists. Additionally, trust in science/scientists moderated the impact of source credibility on liking and sharing intentions: higher trust was associated with stronger intentions to like and share posts from highly credible sources. In summary, these results highlight the importance of general trust in science/scientists and source credibility when communicating information about climate change.

Source Credibility Graphs Trust in Science/Scientist Climate Change Social Media

Talk 6: **Zahra Rahmani** (University of Basel)

Sampling and processing of climate change information and disinformation across three diverse countries

Climate disinformation can undermine public support of climate policies and trust in climate science. Given the limitations of existing preventative measures against disinformation, strategies are needed to increase public exposure to accurate information and decrease exposure to disinformation. Whereas previous research examined the effects of disinformation on climate beliefs and actions, little is known about how people seek out climate-related content (Pro- or Anti-climate scientific consensus) and how this varies between cross-cultural contexts. In a preregistered, sequential information-sampling experiment conducted in the U.S., Germany, and China (Ntotal = 2226), we study how individuals sample and process climate information and disinformation. Participants freely sampled real-world climate related statements, retrieved from social media and validated in previous studies. Across 15 trials, participants decided between two boxes containing Pro-climate or Anti-climate statements, respectively, rating their agreement to the statement and momentary climate concern after each selection. Overall, reading a statement influenced individual climate concern across countries. Initial climate beliefs predicted box choice and this confirmation bias intensified the more information had been sampled. While climate concern was mostly stable, in the U.S., climate concern levels and box choices mutually reinforced each other leading to greater polarization within the sample. The information sampling paradigm offers new perspectives on how people process climate information and navigate information in polarized environments, i.e. distinct information clusters that feature opposing messages and narratives about climate change side by side.

Disinformation climate action climate policies information sampling confirmation bias

Talk 7: **Alexandra Afroditi Asimakopoulou** (University College Dublin)

Exploring the Role of Emotions and Analytical Reasoning in Susceptibility to Food-Related Fake News

Extensive research has shown that memories are inherently unreliable and prone to distortion and decay. Bower’s Network Theory of Affect suggests that memories formed in a particular emotional state are more likely to be retrieved when the individual experiences a congruent emotional state. Additionally, dual-process theory explains that individuals process information through two distinct cognitive systems: a quick, intuitive process (System 1), and a slow, analytic process (System 2). System 1 relies on cognitive heuristics, leading to rapid but often error-prone decisions, whereas System 2 engages in deliberate, analytical reasoning, resulting in more accurate judgments. Research suggests that those who engage in analytic thinking are less susceptible to believing in fake news, and forming false memories for fake news. Furthermore, greater emotional distance increases the likelihood of engaging in analytical thinking. This study investigates how analytical reasoning mediates the relationship between people's emotional state and their susceptibility to fake news. Participants will first undergo an emotional induction, then complete a

Cognitive Reflection Test (CRT) to assess analytical thinking. Finally, they will be presented with 12 emotionally valenced food-related fake news headlines. We expect that people in a positive emotional state will show lower analytical reasoning, leading them to form stronger beliefs and false memories about fake news, while those in a negative emotional state will demonstrate higher analytical reasoning, resulting in fewer false beliefs and false memories. This research sheds light on how emotions and cognitive processes impact memory reconstruction.

Emotional state | Analytical reasoning | False memory | False beliefs | Food-related fake news

Talk 8: **Aleksandra Naganska** (University of Warsaw)

Too good to be true? Financial scarcity and susceptibility to investment scams

Scams are the most common crime in the world, and investment scams are often at the top of the list with financial losses in the billions of USD each year. Disguised as novel guaranteed money-making schemes, revolutionary crypto projects, or limited-time not-to-miss financial opportunities, they tempt individuals by offering exorbitant profits and guaranteed safety. These scams often highlight the benefits and downplay the risks, proving to be too-good-to-be-true offers. One line of research in decision-making draws on scarcity theory. The theory maintains that financial scarcity is associated with a wide range of suboptimal financial decisions. To date, however, scarcity theory has not been applied to the domain of investment fraud - despite it representing a clear case of suboptimal decision. To bridge this gap, the present work was designed to investigate the impact of financial scarcity on response to fraudulent financial investment offers.

To do so, we have designed an online study where we manipulate feelings of financial scarcity. Thereafter, participants (a representative sample of American internet users) are asked to rate legitimate and fraudulent investment offers and indicate their willingness to invest money in these offers. We find that perceptions of risks and benefits are crucial for willingness to invest in scam and legitimate offers. Additionally, we observe relatively high levels of overconfidence among the surveyed sample which points to a greater issue in the general population.

Investment scams | manipulation | risk perceptions | scarcity