Workshop

Content Moderation and Free Speech on Social Media

18th-19th of October 2023 in Munich

Technical University Munich, Richard-Wagner-Str. 1, 80333 Munich, MEETING ROOM B.158

Tuesday, 17 October 2023

Time	Location	Presentations and Activities
19:30-21:00	Löwenbräu–keller	**Pre-workshop meet-up** Pre-workshop meet-up with Bavarian delicacies, drinks and socializing in the Löwenbräukeller (Nymphenburger Str. 2, 80335 München)

Wednesday, 18 October 2023

Time	Location	Presentations and Activities
09:00-09:30	TUM Think Tank	**Registration and Coffee**
09:30-09:45	TUM Think Tank	**Welcome** Welcome by *Yannis Theocharis* (Technical University of Munich) and *Spyros Kosmidis* (University of Oxford)
09:45-11:05	TUM Think Tank	**Session 1** **"Borderline Speech and Networked Governance: TikTok's Moderation Controversies in South and Southeast Asia** *Diyi Liu* (University of Oxford) Content moderation comes with intricate trade-offs and moral dilemmas (Celeste et al., 2023; Jiang et al., 2022), particularly for transnational platforms striving to reconcile consistency with local contextuality. While extensive research has explored the legality and legitimacy of speech governance in democratic contexts (Haggart & Keller, 2021; Suzor et al., 2018), there remains a gap in understanding whether the Western communication order's legitimation tools adequately address the complexities of less-than-democratic

		developing nations (Ramesh et al., 2022; Shahid, 2023). This paper expands the discussion through a case study of how a user-generated content platform originating from China (i.e, TikTok) engages in local speech governance in areas that have been overlooked in scholarship (i.e., South and Southeast Asia). It provides an analytically broad spectrum for understanding different power dynamics that may be said to shape and influence the legitimacy of transnational platform companies in Asia as they become manifest in very different but regionally related contexts. Specifically, the paper posits that power dynamics come to the forefront through "public shocks" -- notable incidents challenging a platform's core principles, prompting operational adaptations (Ananny & Gillespie, 2017:2). To unravel the political economy and power dimensions underlying these algorithmic and regulatory decision-making, I draw on the notion of "networked governance" (Caplan, 2023), which characterises the dynamic legitimation process through interactions among various social actors, who continually rearticulating moderation norms sensitive to each national context. TikTok's meteoric rise has been marked by permanent and temporary bans related to national security, ineffective content moderation, immorality and indecency, and political election interference in SA/SEA. In the paper I summarised the platform's moderation challenges, the legal basis and moral dilemmas at play in these regional controversies, aiming to provide a comprehensive account of contentious or borderline content related to online speech governance in South and Southeast Asia, how they were discussed in the public sphere and how laws and regulations were applied or adapted (or not applied), and the institutional and technical assemblages the platform maintains for borderline moderation. **"Exploring Ideological Biases in Content Moderation on YouTube in the United States"** *Andreu Casas* (Vrije Universiteit Amsterdam) Citizens increasingly rely on social media platforms, such as YouTube, to learn about and engage in politics. The platforms play an active role in content moderation by removing particular posts and accounts. They mostly do so to battle bots and toxic behavior, and so to improve the health of the platforms. However, in recent years conservative and populist groups have argued that the companies also moderate content based on ideology, with an alleged liberal bias. Despite the political and democratic relevance of these claims, empirical analyses exploring potential ideological biases on social media content moderation (or the types or lack thereof) are lacking. In this new project I track a large number (~20k) of YouTube channels that discuss US politics across time – regularly collecting all new content posted by the channels, and checking whether the previously collected content is still active. Then, I use computational methods to model suspensions, at the post and account level, as a function of their ideology and many other confounders.
11:05-11:25	TUM Think Tank	**Coffee break**

11:25-12:45	TUM Think Tank	**Session 2** **"The Politics of Platform Regulation: How Governments Shape Online Content Moderation"** *Robert Gorwa (WZB Berlin)* As digital platforms have become more integral to not just how we live, but also to how we do politics, the rules governing online expression, behaviour, and interaction created by large multinational technology firms --- popularly termed 'content moderation,' 'platform governance,' or 'trust and safety' --- have increasingly become the target of government regulatory efforts seeking to shape them. This book provides a conceptual and empirical analysis of this important and emerging tech policy terrain of 'platform regulation.' How, why, and where exactly is it happening? Why now? And how do we best understand the vast array of strategies being deployed across jurisdictions to tackle this issue? The book outlines three strategies commonly pursued by government actors seeking to combat issues relating to the proliferation of hate speech, disinformation, child abuse imagery, and other forms of harmful content on user-generated content platforms: *persuasive, collaborative,* and *contested* forms of platform regulation. Drawing upon global regulation and public policy scholarship, it then outlines a theoretical model for explaining the adoption of these different strategies across varying political contexts. This model is explored through four qualitative case studies of policy development (Germany, Australia and New Zealand, United States), and is empirically driven by a large number of stakeholder interviews and deliberative policy documents obtained via freedom of information requests. In this talk, I will explore the recent context of content moderation-oriented platform regulation in the United States. I discuss how state-level policy entrepreneurs, bolstered by a rising and revisionist coalition of Conservative anti-tech interest groups, were motivated by Trump's de-platforming to try and shift the platform regulation status quo in the US. Drawing on interviews and new policy documents obtained via FOIA, I examine how these actors used state-level legislation as a way to bypass Federal level gridlock, and worked against significant institutional constraints and industry resistance to enshrine — if only temporarily — a contested platform regulation strategy into Texas and Florida law. **"Citizen Preferences for Online Hate Speech Regulation"** *Simon Munzert* (Hertie School), Richard Traunmüller (University of Mannheim), Pablo Barberá, (University of Southern California), Andrew Guess (Princeton University), JungHwan Yang (University of Illinois at Urbana-Champaign) The shift of public discourse to online platforms, and the resulting widespread visibility of hateful content, have intensified the debate over content moderation by platforms and the regulation of online speech. Designing rules that are met by wide acceptance requires learning about public preferences. We present a visual vignette study combined with a framing and exposure experiment using a sample (N = 2,622) of German and U.S. citizens. To analyze perceptions of online hate speech and preferences for its regulation, we exposed participants to 20,976 synthetic social media vignettes mimicking actual cases of hate speech. We find people's evaluations to be primarily shaped by the type and

		severity of the messages, and less by contextual factors such as the identity of the target or sender. While we find broader support for focused measures like deleting hateful messages, more extreme sanctions like job loss find little support even in cases of extreme hate. We also find evidence for substantial differences in evaluations between gender and ideological subgroups as well as for in-group favoritism among political partisans. Further experimental evidence shows that exposure to hateful speech reduces tolerance of unpopular opinions. Our results provide a fruitful starting point for an evidence-based approach to online content regulation.
12:45-13:45	TUM Think Tank	**Lunch**
13:45-14:45	TUM Think Tank	**Session 3** **"A CERN Model for Studying the Information Environment"** Zoom presentation of a new CERN-like initiative by *Jacob Shapiro* (Princeton University), followed by a discussion within the group. Read more about the initiative here.
14:45-15:40	TUM Think Tank	**Session 4** **"The Future of "Free" Speech: Determinants of Canceling in Academia"** *Nils Weidmann* (University of Konstanz) and *Richard Traunmüller* (University of Mannheim)
15:40-16:00	TUM Think Tank	**Coffee break**
16:00-17:45	TUM Think Tank	**Panel discussion** **"Digital Threats, Content Moderation and Free Speech on Social Media – Perspectives from Politics and Social Media Platforms"**  *Teresa Ott* is Germany's first Hate Speech Officer at the Attorney General's Office Bavarian Ministry of justice. She coordinates and supports the work of special prosecutors at local public prosecutor's offices with regard to the criminal handling of cases involving hate and incitement on the Internet in its various forms. She conducts high-profile investigations herself with her Hate Speech Team at the Munich General Public Prosecutor's Office.

		 *Benjamin Brake* is Benjamin Brake is the head of the newly created "Digital and Data Policy" department at the Federal Ministry of Digital Affairs and Transport. In this position, he reports to Germany's Digital State Secretary Stefan Schnorr. Prior to that, he worked at IBM, where he represented the company's political interests for about 10 years as head of the Berlin office.  *Friedrich Enders* is Government Relations and Public Policy Manager for the DACH region (Germany, Austria and Switzerland) at the entertainment platform TikTok. In addition to political communication, this also includes the development and promotion of partnerships with stakeholders from civil society, science and industry. Prior to the role in the Government Relations team, Friedrich had worked in the Trust and Safety department at TikTok, specifically on community guideline development and enforcement. Previously, he had also worked in politics and consulting.
18:00-19:30	Start Königsplatz	**Social event** **Guided tour – Discovering Munich's Past** A guided tour through (among other things) Germany's Nazi past in Munich. The private tour exclusively for workshop participants starts at the nearby Königsplatz and concludes at Odeonsplatz in Munich. Odeonsplatz is conveniently situated near the restaurant where we'll be heading after the tour.
19:30-21:00	Ratskeller	**Get together and dinner** The dinner will be held at Ratskeller (Marienplatz 8, 80331 München), a renowned institution with a rich heritage dating back to the late 1800s.

Thursday, 19 October 2023

Time	Location	Presentations and Activities
09:00-09:45	TUM Think Tank	**Keynote Molly Roberts and Ruth Appel** **"Partisan conflict over content moderation is more than disagreement about facts"** *Ruth Appel* (Stanford University), Jennifer Pan (Stanford University), and *Margaret Roberts* (University of California) Social media companies have come under increasing pressure to remove misinformation from their platforms, but partisan disagreements over what should be removed have stymied efforts to deal with misinformation in the US. Current explanations for these disagreements center on the "fact gap"—differences in perceptions about what is misinformation. We argue that partisan differences could also be due to "party promotion"—a desire to leave misinformation online that promotes one's own party—or a "preference gap"—differences in internalized preferences about whether misinformation should be removed. Through an experiment where respondents are shown false headlines aligned with their own or the opposing party, we find some evidence of party promotion among Democrats and strong evidence of a preference gap between Democrats and Republicans. Even when Republicans agree that content is false, they are half as likely as Democrats to say that the content should be removed and more than twice as likely to consider removal as censorship.
09:45-11:05	TUM Think Tank	**Session 5** **"Resolving Content Moderation Dilemmas Between Free Speech and Harmful Misinformation."** *Anastasia Kozyreva* (Max Planck Institute for Human Development), Stefan M. Herzog (Max Planck Institute for Human Development), Stephan Lewandowsky (University of Bristol, Bristol, University of Western Australia), Ralph Hertwig (Max Planck Institute for Human Development), Philipp Lorenz-Spreen (Max Planck Institute for Human Development), Mark Leiser (VU-Amsterdam), and Jason Reifler (University of Exeter) In online content moderation, two key values may come into conflict: protecting freedom of expression and preventing harm. Robust rules based in part on how citizens think about these moral dilemmas are necessary to deal with this conflict in a principled way, yet little is known about people's judgments and preferences around content moderation. We examined such moral dilemmas in a conjoint survey experiment where US respondents (N =2,564) indicated whether they would remove problematic social media posts on election denial, antivaccination, Holocaust denial, and climate change denial and whether they would take punitive action against the accounts. Respondents were shown key information about the user and their post as well as the consequences of the misinformation. The majority preferred quashing harmful misinformation over protecting free speech. Respondents were more reluctant to suspend accounts than to remove posts and more likely to do either if the harmful consequences of the misinformation were severe or if sharing it

		was a repeated offense. Features related to the account itself (the person behind the account, their partisanship, and number of followers) had little to no effect on respondents' decisions. Content moderation of harmful misinformation was a partisan issue: Across all four scenarios, Republicans were consistently less willing than Democrats or independents to remove posts or penalize the accounts that posted them. Our results can inform the design of transparent rules for content moderation of harmful misinformation. **"The Politics of Anti-Technology"** *Thomas Zeitzoff* (American University) and Jan Zilinsky (Technical of University Munich) The growth of new technology, in particular new communication technology, has raised questions about technology's role in society. Some point to hate speech, polarization, and radicalization of the electorate, while others have emphasized the democratizing potential of tools that facilitate collective action. Beyond these macro-effects, the individual-level effects of social media on users (ranging from disinformation hazards to harms to mental health) are increasingly coming to the forefront. The coming AI revolution only promises to accelerate these trends. However, people's general attitudes towards new technology and their downstream political consequences remain undertheorized and understudied. Most research focuses on social media and ignores other types of technology like Western medicine. To better understand citizens' reservations about technology, we develop a new anti-technology scale and test it on a diverse sample of Americans. Our scale measures four distinct areas of anti-technology attitudes: 1) attitudes towards social media, 2) attitudes towards artificial intelligence, 3) concerns about modernity, and 4) attitudes towards Western medicine. We show that these anti-tech attitudes map into 3 factors: general anti-tech sentiment, skepticism of technological benefits, and preferences for traditional medicine. We also show that our anti-tech factors are correlated with, but distinct from partisanship, conspiracy attitudes, and individual loneliness. We then show that these anti-tech factors predict attitudes towards technology policies and support for contentious actions against tech companies.
11:05-11:25	TUM Think Tank	**Coffee break**
11:25-12:45	TUM Think Tank	**Session 6** **"Toxic Speech and Limited Demand for Content Moderation on Social Media"** Franziska Pradel (Technical University of Munich), Jan Zilinsky (Technical University of Munich), Spyros Kosmidis (University of Oxford), *Yannis Theocharis* (Technical University of Munich) When is speech on social media toxic enough to warrant content moderation? Platforms impose limits on what can be posted online, but also rely on users' reports of potentially harmful content. Yet, we know little about what users consider inadmissible to public discourse and what measures they wish to see implemented. Building on past work, we conceptualize three variants of toxic

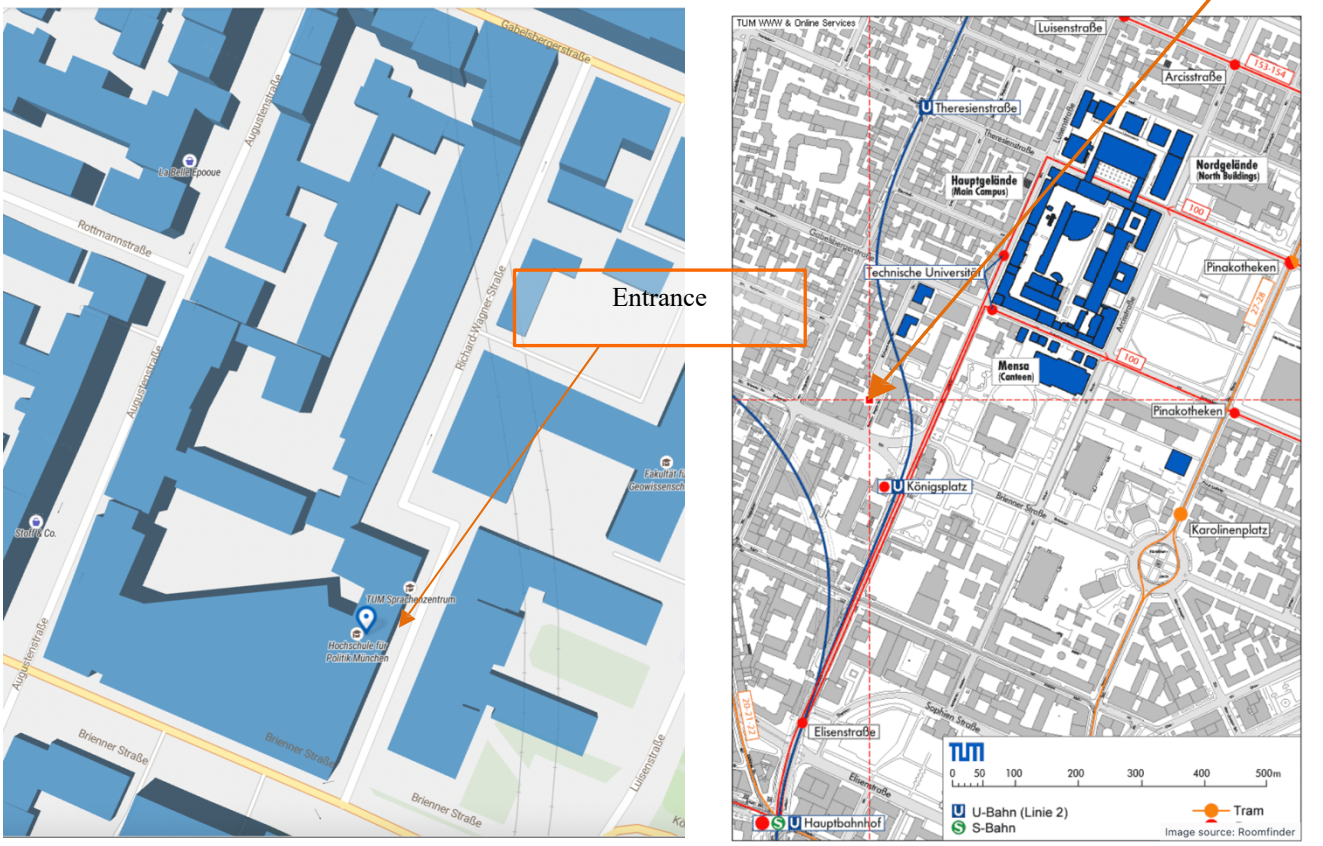
		speech: incivility, intolerance, and violent threats. We present results from two studies with pre-registered randomized experiments (Study 1, N=5130; Study 2, N=3734) to examine how these variants causally affect users' content moderation preferences. We find that while both the severity of toxicity and the target of the attack matter, the demand for content moderation of toxic speech is limited. We discuss implications for the study of toxicity and content moderation as an emerging area of research in political science with critical implications for platforms, policymakers, and democracy more broadly. **"Explaining Preferences for Content Moderation: Tolerance of Toxic Content on Free Speech (and Other) Grounds"** Jan Zilinsky (Technical University of Munich), *Spyros Kosmidis* (University of Oxford), Yannis Theocharis (Technical University of Munich) Perceived barriers to freedom of expression and accusations of politically motivated silencing behaviors ("cancel culture") are a seemingly important cleavage in American politics. But are stated attitudes about freedom of speech attributable to deeply-held values and convictions or are they mere virtue signals? We examine how respondents weigh the value of free speech against the risks from harmful speech and evaluate whether considerations of such trade-offs help explain political attitudes. Next, we explore whether explicit justifications of aggressive speech (e.g., conversational snippets arguing that it is "entertaining" or "normal") move users' preferences for content moderation. Preliminary findings suggests that exposure to excuses for toxic speech may increase support for content moderation.
12:45-14:00	Ella	**Lunch break** Lunch at the close-by Ella (Luisenstraße 33, 80333 München)
14:00-15:20	TUM Think Tank	**Session 7** **"Understanding User Driven Content Moderation"** *Helen Margetts* (Oxford Internet Institute, University of Oxford & Alan Turing Institute), Jonathan Bright (Florence Enoch, Pica Johansson, Francesca Stevens, Alan Turing Institute) User driven online safety technology is an increasingly important part of enabling a safer online environment. This type of content moderation encompasses a wide array of tools offered by platforms to help people tailor their online experiences and protect themselves from harm, for example allowing people to reorganise their newsfeed, specify which kinds of content they would like to see, unfollow or block others, and report content which violates community standards directly to the platform. To some extent these 'user controls' allow people to personalise their online experience. They play a critical role in the upcoming UK Online Safety Bill, which will make it a requirement for large platforms to offer all users accessible and effective safety tools of this kind. However, we know little about the extent to which users are aware of these features, nor the nuanced ways in which they might engage with them.

		This presentation reports on research underway in the Online Safety Team at the Alan Turing Institute, that seeks to enhance our understanding of user driven content moderation. First, we analyse results from a nationally representative survey of 1,160 UK residents to investigate their awareness of, experiences with and attitudes towards seven common user focussed pieces of safety technology currently found on social media platforms. We describe overall awareness and experience, model demographic and attitudinal predictors of engagement, and outline key reasons people give for using different features. Second, we conduct online experiments to deepen our understanding of engagement with one key user control: online reporting mechanisms ('flagging'). Across experiments, we examine how routinely people flag hate speech and abuse, and the extent to which flagging is biased by group identity and political beliefs. This work enhances our understanding of when, how and why people are empowered to protect themselves against online harms. Understanding the social and psychological mechanisms underlying user intervention against such harms is a crucial step towards ensuring a safer online environment. **"Labeling Headlines as AI-Generated Reduces Perceived Accuracy and Sharing Intentions"** Sacha Altay (University of Zurich) and *Fabrizio Gilardi* (University of Zurich) We examine the efficacy of labeling content generated by generative artificial intelligence (AI) systems, such as ChatGPT, as a policy intervention for content moderation on social media platforms. Entities such as the European Commission have asked platforms to put in place technology to label AI-generated to users, drawing parallels to existing measures like the tagging of false news. However, the implications of such labeling remain uncertain; it is conceivable that users may misconstrue the labels, associating AI-generated content with falsehoods. To evaluate these concerns, we conducted a pre-registered survey experiment involving 2,000 U.S. respondents. Our design incorporated four headline categories by varying two elements: truthfulness and authorship (AI vs. human). The study employed five experimental conditions to assess the impact of labeling on content perception: a control group without labels, a group with accurately labeled AI-generated headlines, a group with mixed labeling errors for AI and human-generated headlines, a group with incomplete labeling of AI-generated headlines, and a group where all false headlines were explicitly labeled as such. We measured two primary outcomes: the accuracy ratings assigned to the headlines and the respondents' willingness to share them. Our findings indicate that AI labels lead to lower accuracy ratings and, to a lesser extent, reduced sharing intentions. Specifically, when headlines were marked as AI-generated, respondents were less likely to perceive them as true and were less inclined to share them. These results underscore the complexities involved in designing effective content moderation strategies for AI-generated material.
15:20-15:40	TUM Think Tank	**Coffee break**

15:40-17:00	TUM Think Tank	**Session 8** **"Zones of Contention over Content Moderation"** *Ralph Schroeder* (University of Oxford) This essay provides an overview of some key issues related to the current state of content moderation (CM) on social media platforms. The argument is that there are obstacles to globally applicable rules for CM, and there are both factual reasons for this which also give rise to normative difficulties. The essay proceeds as follows: first, it discusses a few select findings from the growing literature on this topic. Next, it reviews some previous ways of thinking about tackling CM. Then, it gives an overview of the problem globally, distinguishing the 'three effects' in how European, US and Chinese regulatory approaches differ. Thereafter follows an account of how social media platforms should be considered as falling outside of and yet having some similarities to the rules that govern the mediated public arena, which is mainly related to the gatekeeping functions in different types of media systems. Finally, the essay turns to the legitimacy of gatekeeping, and how this legitimacy is unlikely to be globally applicable, but also how, from a normative perspective, such legitimate rule-making for CM should be sought. **"Co-optation and Coercion of Online Influencers: Evidence from Saudi Wikipedia"** *Alexandra Siegel* (University of Colorado Boulder) How do authoritarian regimes use co-optation and coercion of influential internet users to control online information? This paper explores how the Saudi regime co-opted prominent Wikipedia administrators to alter content on sensitive domestic and foreign political topics. I argue that the co-optation and coercion of influential social media users offers regimes an effective tool to manipulate online information environments, with greater plausible deniability and better evasion of content moderation than other forms of computational propaganda. Drawing on a recent ban of Saudi Wikipedia users for coordinated inauthentic activity, I use a two-way fixed effects design and quantitative text analysis of Wikipedia edits to evaluate how banned users' behavior compares to the activity of non-banned users before and after their reported co-optation. I find that Saudi co-optation led to increased editing of pages referencing sensitive political topics, particularly during moments of crisis. This work contributes to our understanding of how authoritarian regimes have adapted longstanding strategies of cooptation, coercion, and information control in the digital age.
17:00-17:15	TUM Think Tank	**Closing** **Recap, closing remarks and discussion** *Yannis Theocharis* (Technical University of Munich) and *Spyros Kosmidis* (University of Oxford)

Workshop Venue

Address:
Hochschule für Politik/Technical University of Munich
Richard-Wagner-Straße 1, 80333 Munich



Contact

Please do not hesitate to contact us if you have any questions via contentmoderation@hfp.tum.de

Prof. Dr. Yannis Theocharis, Department of Governance, School of Social Sciences and Technology, yannis.theocharis@hfp.tum.de

Prof. Dr. Spyros Kosmidis, University of Oxford, Department of Politics and International Relations, spyros.kosmidis@politics.ox.ac.uk

Organized by

Technical University of Munich

UNIVERSITY OF OXFORD

TUM THINK TANK

Reboot Social Media lab