eutema CEO Co-authors Alexander von Humboldt Institute for Internet and Society *AI and Content Moderation* Research Sprint Policy Brief 'Freedom of Expression in the Digital Public Sphere'

Sprint was part of the Global Network of Internet and Society Research Centers research project 'The Ethics of Digitalisation - From Principles to Practices' which aims to foster global dialogue on the ethics of digitalisation among stakeholders from academia, civil society, policy and industry

In contemporary society, much of public discourse and social interaction takes place over online social media platforms, for instance Facebook, YouTube, Reddit or Tik Tok. One of the foundational pillars on which these digital spaces are constructed is freedom of expression. But the line between freedom of expression and hate speech or other challenging content can be ambiguous and difficult to monitor, despite - and because of - the proliferation of algorithmic content moderation (ACM) systems. Adding yet a further layer of complexity is increasing public pressure on privately owned platforms to more soundly police content posted to their sites, and conversely, for legislators to safeguard and uphold freedom of expression through regulation.

It is against this backdrop that the 2020 'AI and Content Moderation' research sprint took place. eutema CEO Erich Prem joined a team of 13 platform governance and research collaborators as part of the Alexander von Humboldt Institute for Internet and Society's (HIIG) 10-week initiative which fell within the framework of the larger Stiftung Mercator-funded Ethics of Digitalisation project. The first product of the sprint is a policy brief, released this month, entitled 'Freedom of Expression in the Digital Public Sphere - Strategies for bridging information and accountability gaps in algorithmic content moderation'.

Currently, social media platforms provide technical infrastructure and tools which facilitate communication and interaction among user communities who engage in public discourse by posting and sharing content, in turn facilitating civic participation in addressing issues of public concern. Despite the dominant position these platforms occupy in society, they remain private entities which set their own rules and guidelines for participation and which employ their own content moderation measures to ensure compliance. Content moderation is the process these platforms use to shape community participation, prevent abuse, and ensure compliance with defined terms of service and standards. It can take many forms, from users reporting posts as violations of specific rules, to automated algorithms flagging potentially illegal content, to banning users. Thus, decisions which affect an individual's ability to participate in the public discourse and which govern user activity and behaviour on platforms are made at the discretion of private companies. It follows that privately administered content moderation can shape and influence the type and extent of discourse taking place. The takeaway then, is that content moderation holds considerable implications for the protection of freedom of expression in the digital public sphere.

Providing further exacerbation is the recent proliferation of ACM systems, which intensify the impact of content moderation on freedom of expression. Platforms use these algorithm-based computational solutions to classify questionable, or to identify specific types of, content in support of moderation. Flagged content can then be used for automated deletion or blocking, or as a pre-moderation support tool for human reviewers. But automated content classification is technically challenging and complex. In its current state, the technology remains context blind, achieves only partial recognition, and produces inaccurate results. In short, it lags far behind humans in recognition rates and capabilities. Yet ACM use is growing for several reasons: posted and shared content volume is surging due to increased users and activity; AI-technology perceived to be faster, and more efficient and cost effective than humans is becoming more readily available; more stringent regulatory frameworks are exerting pressure to monitor and filter illegal content as part of a global shift towards increasing the degree of liability imposed on platforms - including the threat of financial penalties for non-compliance; and, COVID-19 has forced employees from large platforms such as Facebook or YouTube into home office, leaving tasks normally carried out by humans to be instead performed by ACMs.

This increasing reliance on ACM and the resulting effects on the social media ecosystem, combined with concerns about how private decision-making is shaping public discourse and norms by mediating what the public can see, hear and say online, emphasizes that the time has come for governments which support deliberative discussion, self-determination, and inclusive civic participation in the digital public sphere to foster healthy and vigorous public discourse by ensuring increased transparency in platform governance. For this reason, The Freedom of Expression in the Digital Public Sphere policy brief has been developed to inform and support policy makers and legislators. The paper elaborates on privately held digital infrastructure's role in public discourse and the associated - and proliferating - use of ACM systems, with particular focus on their relationship, and risks posed, to freedom of expression. It explains the need for governments to incorporate a more proactive regulatory approach towards the governance of content moderation systems which enable interaction among user communities for public discourse. Finally, it provides a series of proposals and recommendations for safeguarding freedom of expression on online platforms open to public participation.

"I am shocked about the enthusiasm with which algorithms are presented to manage online discourse. The dangers and limitations of this technology are insufficiently discussed although they are clearly visible," said DDr Prem. "We must discuss and develop alternatives before submitting ourselves to an artificial censor agent that platforms can design at their taste."

DDr Prem was part of an interdisciplinary team of experts representing fields including law, engineering, political science and sociology, who co-authored the brief. In addition to this output, two further policy briefs have resulted from the AI and Content Moderation research sprint: Making Audits Meaningful and Disclosure Rules for Algorithmic Content Moderation. More information about the sprint and the Ethics of Digitalisation project can be found on the HIIG Digital Society Blog.

About the AI and Content Moderation research sprint:
The AI and Content Moderation research sprint was a 10-week virtual event hosted by the Alexander von Humboldt Institute for Internet and Society between August and October 2020. It brought together 13 international researchers from a variety of disciplines to tackle challenges presented by automation in content moderation. The result was a series of policy briefs focusing on algorithmic audits and increasing the transparency and accountability of automated content moderation systems.

About the Ethics of Digitalisation - From Principles to Practices:
The Ethics of Digitalisation - From Principles to Practices project aims to develop ground-breaking answers to challenges in the area of conflict between ethics and digitalisation. Innovative scientific formats, research sprints and clinics, form the core of the project; they enable interdisciplinary scientific work on application-, and practice-oriented questions and achieve outputs of high social relevance and impact. The project promotes active exchange at the interface of science, politics and society and thus contributes to a global dialogue on an ethics of digitalisation. Main partners of the project are the Stiftung Mercator, the HIIG, the Berkman Klein Center for Internet and Society at Harvard University, and the Digital Hub Asia.

About the Alexander von Humboldt Institute for Internet and Society (HIIG):
The Alexander von Humboldt Institute for Internet and Society researchers the development of the internet from a societal perspective with the aim to better understand the digitalisation of all spheres of life.

About eutema:
Since 2001, eutema GmbH has been developing strategies in the field of technology development in cooperation with its European partner network. In addition to industry, eutema also advises and supports the public sector. In the field of science and technology policy, the company carries out evaluations of technology initiatives and research programs and provides strategic technology studies.