Personalized AI-based interventions against online norm violations: behavioral effects and ethical implications.

Undesirable online behaviors such as hate speech or cyber-bullying have impacted public discourse, policy makers and platform providers. They can cause substantial harm to individual victims, limit the inclusivity of online settings and shift public discourse and perceived social norms in the online and offline world.

This multi-disciplinary research project combines the expertise from psychology, informatics and ethics (Max Planck Institute for Research on Collective Goods in Bonn and the Department of Informatics at TUM), to address the problem space of how AI-based interventions can be tailored (in personalized ways) and applied to counter online norm violations and to enhance their effectiveness.

The project investigates

1. Whether interventions against online norm violations have differential effects on the attitudinal and behavioral reactions of transgressors, victims and witnesses of those violations depending on their personality and social identity.
2. The ethical ramifications from specific forms of AI-based personalized interventions for the norm transgressor and social media platforms in general.
3. The balance between the deployment of effective personalized interventions, the associated ethical considerations and the need for privacy and data protection, while also considering the applicability of existing technical methods for privacy-preserving AI to learn and deploy personalized interventions.

Can everybody hear me?

2020 activities and results of the project include

- Using a sizable survey experiment, the team found that the perception of group norms predicted attitudinal and behavioral support for counter-speech against homophobic hate. These results suggest that opinion-based identities might become salient in the face of hate speech.
- Using a descriptive approach and a meta-ethical perspective, the team analyzed the impact of de-anonymization on a South-Korean social media platform, Naver. Naver's interventions in response to the South Korea's Network Act and platform-driven action against hate speech had a positive influence in reducing the amount of hate speech. Although they also decreased the total number of comments written overall on the site. Our work suggests that ethical responsibility for online hate speech should not only lie in the hands of the government, but of platforms that host and enable hate speech between citizens.
- Using a comprehensive and systematic review of the technology landscape and literature, the team revealed that moderation on the internet and social media is predominantly reactive, mostly relying exclusively on the deletion of content and banning of users. This identifies a research gap concerning algorithmic moderation for custom communities (such as Facebook's Groups).

Plans for 2021

Based on these initial findings, ongoing work includes systematically iterating the study setup, and manipulate the content of counter-speech to, in particular, match or mismatch participants' political opinions regarding the issue. The aim is to gain a detailed understanding of how the content of counter-speech can match an online user's identity and develop guidelines to maximize interventions' effectiveness in terms of facilitating support for counter-speech and independent counter-speech to subsequent hate. Regarding meta-ethical challenges, the team is currently collaborating on a detailed discussion of transparency, censorship and redress, with the objective of providing forward-looking recommendations for the ethical design of future personalizing moderation tools. Finally, the team aims to develop guidelines on how we can adapt algorithmic moderation technology to assist in context-specific personalized moderation. This work will be presented at the upcoming AAAI Spring Symposium on Implementing AI Ethics (February 2021), and an academic seminar at the University of Magdeburg (January 2021). ●

2020 Paper and Project Highlights

- How Theories of Personal Identity Reveal and Help Clarify Ethical Challenges in Social Media User Modeling (under review) (Engelmann, S., Grossklags, J.)
- Ordinary people as moral heroes or foes? Role model narratives in China's Social Credit System (under review) (Chen, M., Engelmann, S., Grossklags, J)
- A Review of Ethical Considerations for Moderation Tools on Social Media Platforms. (working paper) (Kuo, T., Engelmann, S., Grossklags, J.)
- Naver Closing Pandora's Box: The Effect of Platform Governance on Cyber Harassment on Naver (working paper) (Kang, Nam Gu, Kuo, T., Grossklags, J.)

2020 Conferences

- Online Hate speech and (Automated) Counter Speech, Digital Woche, September 2020
- Personalized AI-based Interventions Against Online Norm Violations, The Responsible AI Forum (TRAIF) Preview 2020, November 2020
- Fighting Hate Speech: Group Norms and Intervention Behavior, Sociology and Psychology Seminar Series, Max Planck Institute for Collective Goods, October 2020
- Fighting Hate Speech: Group Norms and Intervention Behavior, Social Behavior and Decisions Lab Meeting, University of Virginia, November 2020
- AI, Risk Emergence & Zero Trust Networks, International Conference on Complex Systems, July 2020

Principal Investigators

- Prof. Dr. Anna Baumert, TUM School of Education and Max Planck Institute for Research on Collective Goods, Bonn
- Prof. Jens Grossklags, PhD, Department of Informatics, TUM

Researchers

- Dr. Julia Sasse, Senior Research Fellow, Max-Planck-Institute for Research on Collective Goods
- Niklas Cypris, Max-Planck-Institute for Research on Collective Goods
- Severin Engelmann, Department of Informatics, TUM
- Felix Fischer, Department of Informatics, TUM