Introduction to Adversarial Behavior in Collaboration and Social Media Systems Minitrack of the Collaboration Systems and Technologies Track

Christy M.K. Cheung
Department of Finance and Decision Sciences
Hong Kong Baptist University
ccheung@hkbu.edu.hk

Marten Risius
School of Business
University of Queensland
m.risius@business.uq.edu.au

Matthew K.O. Lee
Department of Information Systems
City University of Hong Kong
cbmatlee@cityu.edu.hk

Christian Wagner
School of Creative Media
City University of Hong Kong
c.wagner@cityu.edu.hk

Social media platforms facilitate coordinated adversary behaviors (e.g., coordinated inauthentic behavior, cybermobbing, distortion of the public discourse, proliferation of unreliable information, relinquished on- and offline privacy). Digital technologies enable transferring traditional adversarial behaviors into the online environment (e.g., bullying, harassment, disinformation) or generate new adversarial behaviors altogether (e.g., doxing).

Addressing the unique challenges of adversarial behavior in collaboration and social media systems requires a focus on the technological implications of online adversarial behavior. This minitrack offers a forum for research ideas that consider the interaction between digital technologies and adversarial behavior to understand the role of digital technologies in enabling (Paper 1) or countering adversarial behaviors online (Paper 2). This year, two papers were selected for inclusion in the proceedings.

The first paper, titled "Are Deep Learning-Generated Social Media Profiles Indistinguishable from Real Profiles?" by Sippo Rossi, Youngjin Kwon, Odd Harald Auglend, Raghava Rao Mukkamala, Matti Rossi, and Jason Thatcher, address the issue of fake social media profiles. The authors report an experiment with 375 participants to investigate social media users' ability to discern fake and real profiles. To that end, they deploy deep learning-algorithms and text generators to create fake profiles and posts. When asked to report the suspiciousness of components, users demonstrate no significant differences in their ability to distinguish fake and real pictures, tweets, names, or twitter handles. This study alludes to the growing threat that state-of-the-practice algorithms pose for improvident social media users.

The second paper, "Toward Designing Effective Warning Labels for Health Misinformation on Social Media" by Huma Varzgani, Nima Kordzadeh, and Kyumin Lee, takes a first step towards mitigating the issue of health misinformation on social media. To that end, the authors propose different warning label designs that differ regarding the background color warmth, the level of abstractness in the warning message, and the assertiveness of the text. They relate these misinformation warning message characteristics to different protective behaviors against misinformation threats (i.e., verifying labeled content, avoid using labeled content, avoid sharing labeled content). Following and testing the proposed designs will help develop empirically guided, persuasive warning labels to counter disinformation.

We thank the authors for submitting their wonderful works to this new minitrack. Their attempts help to deepen and broaden our understanding of the impact of social media. In particular, how information technology can support and mitigate adversarial behavior on social media.

- This work was supported by a fellowship award from the Research Grants Council of the Hong Kong Special Administrative Region, China [HKBU SRFS2021-2H03].
- Marten Risius is the recipient of an Australian Research Council Australian Discovery Early Career Award (project number DE220101597) funded by the Australian Government.

URI: https://hdl.handle.net/10125/102644
978-0-9981331-6-4
(CC BY-NC-ND 4.0)