Zachary Yang

McGill University
Montreal, Quebec
Canada
H3A 0G4

zachary.yang@mail.mcgill.ca
https://rstzzz.github.io/

1 Research interests

Ensuring safe online environments is a formidable challenge, but nonetheless an important one as people are now chronically online. The increasing online presence of people paired with the prevalence of harmful content such as toxicity, hate speech, misinformation and disinformation across various social media platforms (Ciftci et al., 2017; Watanabe et al., 2018; Mohan et al., 2017; Döring and Mohseni, 2020) and within different video games (Silva et al., 2020) calls for stronger detection and prevention methods. According to the Anti-Defamation League's 2023 report, toxicity in gaming is "now so pervasive that it has become the norm for many players" (ADL, 2023). Moreover, concerns among experts are rising about the potential for advanced AI to cause significant harm through manipulation, even before ChatGPT. Sophisticated AI-assisted information operations have already emerged as a growing concern (Hitkul et al., 2021; McKay and Tenove, 2021; Tucker et al., 2017). Already in 2022, systems like Cicero, an AI language agent, had demonstrated capabilities in persuasion and deception within social gaming environments (FAIR et al., 2022).

To foster a healthier online community, companies have experimented with various approaches to curb the dissemination of toxic and harmful content. These efforts range from word censorship and player bans to content moderation and flagging controversial posts for review.

My research interests primarily lie in **applied natural language processing for social good**. Previously, I focused on measuring partisan polarization on social media during the COVID-19 pandemic and its societal impacts (Yang et al., 2021, 2024b). Currently, at Ubisoft La Forge, I am dedicated to **enhancing player safety within in-game chat systems** by developing methods to **detect toxicity** (Yang et al., 2023), **evaluating the biases** in these detection systems (Van Dorpe et al., 2023), and **assessing the current ecological** state of online interactions (Yang et al., 2024a). Additionally, I am engaged in **simulating social media environments using LLMs** to ethically test detection methods, **evaluate** the effectiveness of current mitigation strategies, and potentially introduce new, successful strategies.

1.1 Safety With In-Game Chat Systems

Ensuring player safety in online games begins with effectively **detecting and preventing toxicity within in-game chat systems**. As more games feature online multiplayer modes with team and all-chat options, players engage in conversations through both text and speech. While definitions of toxicity and hate speech vary among researchers and industry platforms, we adhere to the definitions outlined by the Fair Play Alliance (Lewington, 2021). My initial focus was on improving the detection of toxicity (Yang et al., 2023). Previous research primarily focused on social media, revealing that incorporating the context of parent posts did not enhance performance. Since in-game conversations are more cohesive, I integrated techniques from dialogue systems, including previous chat lines and speaker segmentation, to model multi-turn conversations. This enabled the creation of a context-aware model capable of detecting toxicity in real-time game chat.

While advancing these LLMs to detect toxicity is crucial, addressing the potential biases inherent in them is equally important. Consequently, our team **measured identity biases** using a game-focused dataset (Van Dorpe et al., 2023). Inspired by reactivity analysis, we had users annotate whether a sentence was toxic. We generated sentences typical of in-game chat while replacing key words with specific attributes (e.g., black, trans), groups (e.g., white, young people, women), and personas (e.g., artist, streamer). This approach allowed us to measure whether detection algorithms reacted differently to certain terms, leading to unfair treatment of specific groups of players, either through over-penalization or under-penalization.

To fully grasp the current state of toxicity within in-game chat systems, we ran our detection system on a full year's worth of chat data (Yang et al., 2024a). This research examined in-game events, the number of players and matches played, and the types of games. We recognize that any *deployed system will naturally elicit reactions from players*. A holistic approach that considers both the technical aspects of toxicity detection and the socio-cultural environment of online gaming commu-

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 64–67

nities is essential. By gaining a **comprehensive** understanding of these factors, the player safety team can devise more effective strategies to foster a **safer and more inclusive gaming ecosystem**. Capturing the current ecological state before deployment allows us to measure the impact of this detection system in conjunction with any mitigation strategies deployed.

1.2 Simulating Social Media w/ LLMs

With the rise of generative LLMs, the question arises whether they can be utilized to **simulate high-fidelity reflections of social environments**, creating a sandbox mode that allows us to **ethically test detection and mitigation strategies** for social harms such as manipulation during election discourse, the spread of toxicity, hate speech, misinformation, and disinformation.

LLMs have demonstrated the ability to reflect political attitudes (Argyle et al., 2023), showcase personality traits (Serapio-García et al., 2023), and simulate social interaction (El-Kishky et al., 2022; Törnberg et al., 2023). Researchers have already begun using LLMs on a small scale, such as simulating a small town (Park et al., 2023) and social media (Törnberg et al., 2023). Our current work at my research lab focuses on expanding these simulations to a larger scale using the open-source social media platform Mastodon as the environment. We will attempt to employ personas that reflect reality-matching demographics and activity/network attributes from massive Twitter datasets (Pelrine et al., 2023b; Yang et al., 2021; Pelrine et al., 2023a; Orlovskiy et al., 2024). Additionally, we will then introduce benign agents fine-tuned for varying levels of susceptibility to misinformation, mirroring human populations (Liu et al., 2023), and malicious agents that would replicate severe manipulation threats. This controlled setting will then enable us to quantify manipulation effects and assess the effectiveness of proposed defenses, yielding broad applications across AI safety, social science, and policy.

2 Spoken dialogue system (SDS) research

The research on player safety systems is closely connected to spoken dialogue system research, as players frequently communicate through text and speech. Leveraging LLMs to simulate these social environments allows us to ethically test current prevention methods and understand the effectiveness and potential unintended consequences of various mitigation strategies. Spoken dialogue systems, such as chatbots and virtual assistants, rely on natural language processing and generation, which are also fundamental to LLMs. By studying the manipulation and mitigation of social harms in these simulations, we can develop more robust and ethical dialogue systems capable of detecting and preventing misinformation, hate speech, and other malicious content in real-time interactions. This cross-disciplinary approach enhances the safety and trustworthiness of AI-driven communication technologies in both written and spoken forms, ultimately contributing to a more secure and inclusive digital environment.

3 Suggested topics for discussion

- Understanding and mitigating social harms: Addressing toxicity and misinformation through high-fidelity simulation environments.
- Enhancing safety in online environments: multimodal models, handling multi-lingual conversations (where a sentence can contain more than one language), and addressing accents and region-specific dialogue.
- Personification of LLM agents: Developing coherent responses based on backstory and personality.
- Ethically simulating social media sandbox environments at scale with LLM agents: Including the posting of text, speech, images, and video.
- Re-balancing the playing field between good and bad actors: Strategies for countering societal-scale manipulation.

References

ADL. 2023. Hate is no game: Hate and harassment in online games 2023. *Anti-Defamation League* https://www.adl.org/resources/report/hate-no-game-hate-and-harassment-online-games-2023.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis* 31(3):337–351.

Tuba Ciftci, Liridona Gashi, René Hoffmann, David Bahr, Aylin Ilhan, and Kaja Fietkiewicz. 2017. Hate speech on facebook. Fourth European Conference on Social Media Research, pages 425–433.

Nicola Döring and M. Mohseni. 2020. Gendered hate speech in youtube and younow comments: Results of two content analyses. *Studies in Communication and Media* 9:62–88. https://doi.org/10.5771/2192-4007-2020-1-62.

Ahmed El-Kishky, Thomas Markovich, Serim Park, Chetan Verma, Baekjin Kim, Ramy Eskander, Yury Malkov, Frank Portman, Sofía Samaniego, Ying Xiao, and Aria Haghighi. 2022. Twhin: Embedding the twitter heterogeneous information network for personalized recommendation. In *Proceedings*

of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. ACM, KDD '22. https://doi.org/10.1145/3534678.3539080.

Meta Fundamental AI Research Diplomacy Team FAIR, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. 2022. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science* 378(6624):1067–1074.

Hitkul, Avinash Prabhu, Dipanwita Guhathakurta, Jivitesh jain, Mallika Subramanian, Manvith Reddy, Shradha Sehgal, Tanvi Karandikar, Amogh Gulati, Udit Arora, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2021. Capitol (pat)riots: A comparative study of twitter and parler.

Robert Lewington. 2021. Being 'targeted' about content moderation:. *Fair Play Alliance* pages 1–21. https://fairplayalliance.org/wp-content/uploads/2022/06/FPA-Being-Targeted-about-Content-Moderation.pdf.

Yanchen Liu, Mingyu Derek Ma, Wenna Qin, Azure Zhou, Jiaao Chen, Weiyan Shi, Wei Wang, and Diyi Yang. 2023. From scroll to misbelief: Modeling the unobservable susceptibility to misinformation on social media. *arXiv preprint arXiv:2311.09630* .

Spencer McKay and Chris Tenove. 2021. Disinformation as a threat to deliberative democracy. *Political research quarterly* 74(3):703–717.

Shruthi Mohan, Apala Guha, Michael Harris, Fred Popowich, Ashley Schuster, and Chris Priebe. 2017. The impact of toxic language on the health of reddit communities. Canadian Conference on Artificial Intelligence, pages 51–56. https://doi.org/10.1007/978-3-319-57351-9_6.

Yury Orlovskiy, Camille Thibault, Anne Imouza, Jean-François Godbout, Reihaneh Rabbany, and Kellin Pelrine. 2024. Uncertainty resolution in misinformation detection. https://arxiv.org/abs/2401.01197.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. pages 1–22.

Kellin Pelrine, Anne Imouza, Camille Thibault, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph, Jean-François Godbout, and Reihaneh Rabbany. 2023a. Towards reliable misinformation mitigation: Generalization, uncertainty, and gpt-4. *arXiv preprint arXiv:2305.14928* .

Kellin Pelrine, Anne Imouza, Zachary Yang, Jacob-Junqi Tian, Sacha Lévy, Gabrielle Desrosiers-Brisebois, Aarash Feizi, Cécile Amadoro, André Blais, Jean-François Godbout, et al. 2023b. Party prediction for twitter. *arXiv preprint arXiv:2308.13699* .

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models.

Bruno Silva, Mirian Tavares, Filipa Cerol, Susana Silva, Paulo Alves, and Beatriz Isca. 2020. Playing against hate speech -how teens see hate speech in video games and online gaming communities. *Journal of Digital Media and Interaction* 3:34–52. https://doi.org/https://doi.org/10.34624/jdmi.v3i6.15064.

Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. 2023. Simulating social media using large language models to evaluate alternative news feed algorithms. *arXiv preprint arXiv:2310.05984* .

Joshua A Tucker, Yannis Theocharis, Margaret E Roberts, and Pablo Barberá. 2017. From liberation to turmoil: Social media and democracy. *J. Democracy* 28:46.

Josiane Van Dorpe, Zachary Yang, Nicolas Grenon-Godbout, and Grégoire Winterstein. 2023. Unveiling identity biases in toxicity detection : A game-focused dataset and reactivity analysis approach. In Mingxuan Wang and Imed Zitouni, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*. Association for Computational Linguistics, Singapore, pages 263–274. https://doi.org/10.18653/v1/2023.emnlp-industry.26.

Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki. 2018. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access* 6:13825–13835. https://doi.org/10.1109/ACCESS.2018.2806394.

Zachary Yang, Nicolas Grenon-Godbout, and Reihaneh Rabbany. 2023. Towards detecting contextual real-time toxicity for in-game chat. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, pages 9894–9906. https://doi.org/10.18653/v1/2023.findings-emnlp.663.

Zachary Yang, Nicolas Grenon-Godbout, and Reihaneh Rabbany. 2024a. Game on, hate off: A study of toxicity in online multiplayer environments. *ACM Games* Just Accepted. https://doi.org/10.1145/3675805.

Zachary Yang, Anne Imouza, Kellin Pelrine, Sacha Lévy, Jiewen Liu, Gabrielle Desrosiers-Brisebois,

Jean-François Godbout, André Blais, and Reihaneh Rabbany. 2021. Online partisan polarization of covid-19. In *2021 International Conference on Data Mining Workshops (ICDMW)*. IEEE, pages 893–901.

Zachary Yang, Anne Imouza, Maximilian Puelma Touzel, Cecile Amadoro, Gabrielle Desrosiers-Brisebois, Kellin Pelrine, Sacha Levy, Jean-Francois Godbout, and Reihaneh Rabbany. 2024b. Regional and temporal patterns of partisan polarization during the covid-19 pandemic in the united states and canada. https://arxiv.org/abs/2407.02807.

Biographical sketch

Zachary Yang is a PhD candidate at McGill University, supervised by Professor Reihaneh Rabbany, specializing in applied natural language processing for social good. His research focuses on studying toxicity, misinformation, and polarization in games and social media. Zachary is also a member of Mila - Quebec AI Institute and the Centre for the Study of Democratic Citizenship.

Previously, he developed scalable methods to measure partisan polarization on social media during the COVID-19 pandemic, with his work published in IEEE VIS and ICDMW. Currently, his research at Ubisoft La Forge aims to improve and prevent toxicity detection within game chat and create industry-leading player content safety systems. This work has led to publications in EMNLP 2023 and a presentation at the Ethical Games Conference in 2024. After completing his PhD, Zachary plans to join a research lab in industry, bridging academia and industry to deploy more systems with humans-in-the-loop, enhancing safety and ease of use.