Georgii Zherebilov
Universitat Pompeu Fabra, Department of Economics and Business
C/ Ramon Trias Fargas 25–27, Barcelona, Spain
+34 744 649 299
georgii.zherebilov@upf.edu

I am a 5th-year PhD student in Economics at Pompeu Fabra University, Spain. I completed two years of PhD coursework at the Barcelona School of Economics and Pompeu Fabra University, as well as a 2-year master's in Economics degree from EIEF and LUISS, Rome. I am currently working on my dissertation under the supervision of Ruben Enikolopov and Maria Petrova. I expect to complete the degree by September 2027. My research interests lie in Applied Microeconomics and Political Economy, with a focus on digital platforms and political communication. Across my dissertation projects, I study how social media algorithms, automated accounts, and AI tools shape information environments, user behavior, and the diffusion of political content. My work combines large-scale online data, causal inference, and modern computational tools, with the goal of generating policy-relevant insights on the governance of digital communication and the use of AI to fight misinformation.

As part of my dissertation, I'm currently working on three projects. My first project examines how the introduction of AI-written fact-checking notes affected the X Community Notes platform, a crowdsourced fact-checking tool launched in 2021 on Twitter (now X). Community Notes allows users to suggest short corrective notes on posts they view as misleading; notes that receive sufficient agreement from community evaluators are displayed publicly. Prior work finds that Community Notes often intervene too late in the misinformation diffusion process (Chuai et al., 2024; Slaughter et al., 2025), in part because political partisanship slows down agreement (Allen et al., 2022). AI-generated notes may help mitigate these frictions by lowering writing costs and producing less polarized content.

In the first part of the project, I exploit the pilot launch of AI-written notes, during which accepted AI notes were not labeled as AI-written when displayed under tweets. With the full launch of AI notes, the labels that show AI authorship were introduced. This disclosure allows me to employ a difference-in-differences design comparing misleading tweets with AI-written and human-written notes before and after the disclosure change. This natural experiment allows me to causally identify how users perceive AI as a fact-checker relative to humans—specifically, whether revealing AI authorship affects the trustworthiness and persuasive power of fact-checks, as reflected in engagement with misinformation.

In the second part, I study how AI authorship affects voting outcomes within the Community Notes system. Using a randomized experiment, I generate notes with an AI writer and randomly assign them to be posted either from an AI-annotator account—making authorship salient—or from a human-annotator account. This design allows me to isolate the effect of perceived AI authorship on note credibility, voting behavior, and measures of note quality. Together, the two parts provide evidence on how AI tools can be incorporated into platform-level fact-checking and whether both regular users and community fact-checkers trust AI-generated fact-checks across a broad range of misleading content

In a second project, joint with Alexey Makarin, Andrey Simonov, Rafael Jiménez-Durán, and Sergey Sanovich, we study how coordinated bot activity shapes political communication on social media. We partner with Botnadzor, a Russian grassroots organization that develops algorithms to detect bot accounts, and use a unique dataset containing comments from over 100,000 identified bots across thousands of VK groups. Using platform-level variation from mass bot-ban waves and predictable bot working hours, we estimate how bots affect engagement with posts, the tone and structure of discussions,

and the political sentiment of commentators. To identify downstream effects of bot exposure on political attitudes and online behavior, we run a randomized experiment in which participants install a browser extension that highlights bot comments and tracks browsing behavior. This project provides evidence on how coordinated bot accounts influence political discourse and user beliefs online.

Finally, in a joint project with Juan S. Morales and Annalí Casanueva Artís, we study social media narratives surrounding the BLM movement in the aftermath of George Floyd's death. We apply LLM-based methods to classify actors and sentiment in the universe of BLM-related tweets produced in the month following the killing. Linking geocoded tweets to a comprehensive database of hate-crime reports, we examine how narratives about police and race affect citizens' likelihood of reporting hate crimes and local agencies' decisions to publish reports. This project aims to shed light on how shifts in social media narratives about policing and race can translate into real-world outcomes, influencing hate-crime incidence, reporting behavior, and institutional transparency.

References:

Chuai, Y., Tian, H., Pröllochs, N., & Lenzini, G. (2024). *Did the roll-out of Community Notes reduce engagement with misinformation on X/Twitter?* Proceedings of the ACM on Human-Computer Interaction, 8(CSCW2), 1–52.

Slaughter, I., Peytavin, A., Ugander, J., & Saveski, M. (2025). *Community Notes reduce engagement with and diffusion of false information online.* Proceedings of the National Academy of Sciences, 122(38), e2503413122.

Allen, J., Martel, C., & Rand, D. G. (2022). *Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program.* In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (pp. 1–19).