Research Statement: Fostering Digital Trust

Swapneel Mehta *Boston University and MIT*

My career goal is to conduct research and develop responsible computing systems that foster digital trust and credibility, identifying the social, technological, and economic drivers of erosion of trust on digital platforms. Over the last two decades, we have witnessed a historic, yet consistent, degradation of trust in mass media and democratic institutions[1]. The role of media as a bastion of truth has been significantly weakened and public discourse polluted with a barrage of misleading claims and conspiracies. Even for experts it is difficult to tell with certainty whether we can trust what we see online. Access to practically costless means of producing synthetic media with generative artificial intelligence (GenAI) has only exacerbated this issue, resulting in polarization and social divide. **How do we reverse the consequences of the erosion of digital trust?** I contend that we need to study longitudinal user behavior, understand the consequences of online harm, evaluate interventions to protect users, and employ them to design responsible computing systems. To address misinformation, we must reorient platform incentives to penalize the production of false claims instead of placing the onus on platforms and end users to counter their influence. I discuss a selection of my work in the **social and data sciences** to highlight how I intend to pursue this research agenda.

1 Social Drivers: Proliferation of Misleading Claims on Platforms

Social platforms are information highways connecting the most far-flung regions of the world in milliseconds. Every single minute, there are 2500 videos uploaded to YouTube, over 1.3 million calls made over WhatsApp, 4 million likes generated on Facebook, and nearly a million images shared on Instagram. With the Sunday Times, my nonprofit team at SimPPL built a platform[2] to study the sharing of 1,000+ articles from Russia Today and Sputnik News on Twitter (now, X), focusing on coordinated link-sharing and gathering 14M+ accounts and 70M+ tweets that resulted in Twitter's Site Integrity Team reaching out for us to report 400,000 accounts to them. We recognized the risks from the rise in large language model (LLM) driven bot accounts as we designed behaviorally-aware systems to detect them using machine learning (ML) [9]. To understand their influence, we conducted a 3-part investigation into 600 groups and pages on Meta, measuring the reach of gendered harassment and political targeting campaigns to 95M+ users. It received national news coverage[3] and was the subject of Meta's Adversarial Threats takedown in Q1, 2024 [3]. I noticed that while mainstream platforms are much better studied and understood, there is an emerging ecosystem of alternative social platforms that wield a powerful influence over politics but are barely studied in comparison. I wanted to gauge how interactions with low credibility information vary on unmoderated or lightly moderated platforms. To do this, my team and I monitor the spread of debunked disinformation on Telegram, identifying over 4,500 channels that engaged in the cross-sharing of URLs from debunked articles [2]. We noticed that when new events occurred on the ground such as the Moscow Blasts in 2024, a subnetwork of channels were used to launch information operations. Extending our graph-theoretic analysis, we worked with former intelligence agency consultants to semantically organize and cluster channels, identifying "bridge" channels that serve as hubs for the amplification of content across smaller cliques of channels, creating a viable node for real-world interventions to counter the propaganda machinery.

In a similar vein, modern influencers often serve as a hub for the amplification of content to a large (4-5M monthly active users), ideologically congruent audience on platforms like Truth Social. We identify the statistically significant social drivers of influence in a first-of-its-kind longitudinal study [6] of articles from 1500+ media websites shared in 2M+ *truths* (posts) across 2 years. We find a causal effect underlying an influencer sharing a news article that increases the *re-sharing* of that news article by audiences (via a new post containing the same link); also finding that spammers succeed at gaining influence on the platform in the long term. In related work into community-level information-sharing on Truth Social, we also found, contrary to a commonly held opinion, that there exist communities of users engaged in rational discourse,

[1] https://www.statista.com/statistics/471240/perceived-objectivity-mass-media-usa

[2] https://parrot.report

[3] https://en.prothomalo.com/bangladesh/politics/cene485ilu

sharing Wikipedia articles on the platform [15]. Our work underscores a need to take a nuanced, policy-oriented view of how users behave in right-leaning communities, reflecting similar ideas from [7]. I pitched our work competing with Visa, MassMutual, Workday, Duolingo, etc. at MIT and was successfully selected to lead a capstone team at their Analytics Lab led by Prof. Sinan Aral, aiming to conduct a complementary, longitudinal assessment of community formation, news-sharing, and engagement on Truth Social. Beyond traditional social platforms, it is important to evaluate channels like radio and broadcast television that result in exposure and offline misinformation sharing in social settings, undetectable without explicit means of measurement. In a novel application, my team of capstone mentees at NYU Data Science is working with the United Nations Peacekeeping Operations to advance their search over radio (audio) data from East Africa to identify relevant political narratives; for instance President Joe Biden's failing health ahead of the 2024 U.S. Presidential elections, shown in a demo we have created for the UN Team actively collaborating with us[4].

2 Technological Drivers and Interventions to Mitigate Misleading Information

Spam, hate, fraud, and fake ads are widely prevalent on all manner of digital platforms and it is only getting worse with AI-powered content discovery[5]. Content recommendation algorithms are an opaque distribution mechanism that is not easily possible to intervene on. As an example of why this is the case, Twitter had multiple regional teams applying context-specific rules to prevent the gaming of these systems by local actors, akin to a 'bandage' applied to treat their concern. With multiple bandages from tens of teams, applied across 15 years of their growth, it was impossible to tell what rule set or behavioral change may have caused a harm to pass under their radar. I spent time designing, scaling, and deploying a multimodal recommendation system that attached *trending* hashtags to content creator videos, in order to improve their discoverability online, for a now-patented algorithm at Adobe [12]. Intending to demonstrate the need for systems robust to being manipulated by adversarial actors, I reimplemented Reddit's ranking algorithms and demonstrated how it could be gamed for misinformation amplification using real-world data[13], leading to the design of a multiagent simulator that I am extending.

Platforms deploy interventions as preventative measures that are intended to advance user safety and platform health by limiting harm. In a prominent example of Twitter's interventions in the form of warning labels applied to former President Donald Trump's tweets around Nov. 2020, we causally identified a Streisand effect in the form of increased favorites received by these tweets on the platform. Interestingly, there was no corresponding increase identified in the sharing of these tweets via posts on Facebook, Reddit, and Instagram, that otherwise would link back to Trump's tweets [10]. This suggests that the labeling intervention did have its intended effect of limiting the visibility of misleading election-related claims, but only had it *off-platform*. It led to invited talks and my internship at Twitter where I built a state-of-the-art early (civic) misinformation classifier. Deliberate interventions of this nature have a natural complement in external stimuli that periodically affect platform users, such as outages which cause 'downtime', or the unavailability of platforms. We tracked critical outages that result in the lack of short-term feedback, filtering to posts containing either low or high-quality news URLs [11]. We discovered that in the short term (1-2 weeks), repeated outages increase the volume of low-quality news-sharing on Reddit, but over a longer period (8 weeks), they significantly decrease low-quality news sharing, while increasing the high-quality news posted on the platform, advancing insights we identified that relate with sharing false, but *interesting* information online [14]. These findings underscore the need to complement interventions research with psychosocial experiments to examine the incentives that drive information sharing behavior in an attention economy.

[4] https://drive.google.com/file/d/1ajEt7gZ2wW-MjeHw90EiRbiDyPDMFU3O/view?usp=sharing

[5] https://www.bleepingcomputer.com/news/google/googles-new-ai-search-results-promotes-sites-pushing-malware-scams

3 Economic Drivers: Production of Consumer Harms

Social networks, video streaming services, and e-commerce platforms are designed as two-sided marketplaces serving user-generated information from producers to consumers. Producers demand their First Amendment-protected right to free speech, while consumers demand the right to selectively filter what they consume. However, when either of these rights is abused, the other suffers significantly; producers could create spam, hate, and fake news causing consumer harm, whereas consumers could select into filter bubbles resulting in highly polarized audiences. Because existing interventions place the burden of mitigating harm on either the platform or the consumer, we allow unfettered production of false claims, each causing externalities that increase this burden, until market collapse, because markets do not self-correct externalities. Based on economic theories from Coase and Stiglitz [16], we propose that in digital markets, the harms arising from misinformation production be measured via the magnitude of decision error that they induce on the consumer's part. The shift to measuring decision error internalizes the market externality and allows for an intervention to achieve equilibrium between producers and consumers in the marketplace. In our intervention, we associate a relative economic cost to the production of misleading claims; for instance advertisers may optionally choose to escrow an amount *per user*, called a 'truth warrant', in order to signal the veracity of their claims. If users find the advertised claim to be misleading, they can challenge the claim and post a fraction of the escrowed amount, to potentially win the full escrowed amount if a peer jury (think, *X's Community Notes*) adjudicates in their favor [19]. If not, the advertiser retains their escrowed amount and the challenging user loses theirs. We built our own interactive two-sided e-commerce marketplace and ran an experiment with human buyers purchasing products over 1350 rounds of advertisements by pre-defined honest and cheating bot sellers. We found that our intervention not only increased profits for honest advertisers but also penalized dishonest advertisers [5, 4]. In fact, since Amazon introduced GenAI agents for sellers, we tested agentic sellers using large language models (LLMs) and found that while they strategize to maximize product sales dishonestly in the control market, the intervention significantly reduces the profits deceptive strategies obtain [1]. We successfully found the emergence of a well-defined set of strategies that the agentic sellers relied on, reflecting a rational set of decisions. We are actively exploring how LLMs might behave as decision-support systems by training them on Community Notes and engaging them in multiple rounds of debates with early results stating that they could reach a consensus that is 88% accurate in mimicking human note writers [8]. We are in the middle of deploying field experiments in two-sided marketplaces involving human consumers and human producers to compare how their strategies will be different from their responses to those of adaptive and agentic bot participants in the market.

4 Future Work

There are ways in which others have attempted to improve trust in online information but very few recognize the need for community-based participatory design processes and their role in the eventual efficacy of interventions [18]. I aim to pursue human-centered design approaches in pursuing research to improve the safety of online users and mitigate harms caused online. Through my interdisciplinary work, I have access to a unique set of collaborators working with whom helps to accelerate progress in both fields. This started with AI for the natural sciences at CERN, AI for journalism at SimPPL, extended to behavioral economics in my postdoc, and now healthcare for the GenAI systems we are deploying in India and Bangladesh. I would like to advance my research focused on **estimating the causal effects of cross-platform interventions** and **redesigning platform incentives** to rebuild digital trust.

I continue to work with Questrom's Platform Governance Lab[6] that I helped set up and grow (particularly with Marshall van Alstyne and Nina Mazar), where I have mentored over 15 undergraduates this year. In the short term, I will continue working on social and technological drivers to understand the dynamics of influence, role of algorithms on online trust, and field experiments with information auditing. As a first project, we have hourly snapshots of engagement with each post from 30 top political leaders on Truth Social for weeks before, during, and after President Joe Biden resigned as the Democratic Presidential nominee and as Kamala Harris took the lead all the way through to her loss–this is data that can *never be recollected* without platform support, making it incredibly valuable to discern patterns in civic engagement

[6] `https://truthmarket.com/people`

on electoral issues [7]. Second, at SimPPL, we have already aggregated half a billion posts (and counting) from 6+ social platforms with and without platform API support and are developing a cross-platform (Reddit, Meta, Instagram, Bluesky, Truth Social, Telegram, Moj, Bitchute) semantic search pipeline to study how narratives propagate across communities in real-time. In the next year, we are scaling our research infrastructure to unprecedented speeds and volume via a platform we call Arbiter (being piloted with fact-checkers in Mongolia), that would allow similar longitudinal analysis across multiple social platforms, including multimodal analysis. Third, extending work on the effects of content creators and political elites, I want to understand how influence drives offline behaviors. For this, I am collaborating with India's largest national newsrooms examining how influencers affect the public.

In the longer-term, I will advance work on economic incentive structures for emergent social platforms, identify strategies for policy teams to contribute to harms reduction, and devise collaborations to advance platform governance centered around advertising as a business model. This emerges from my postdoctoral research on limiting misleading advertisements in digital markets[8] [5].

4.1 Tech and Policy Partnerships to Strengthen the Information Ecosystem

Social media platforms are well aware of the regulatory consequences of allowing online harms to affect their users, and invest in developing civic integrity and trust and safety teams. However, the challenge is these teams are siloed, treated as a 'cost center', and end up having limited impact in many contexts, until harms emerge. Platforms seldom coordinate their intervention efforts with the notable exception of child safety and illegal content where coalitions are established for collaborative action. This leaves the field open to external researchers to independently or in platform-collaborations (Social Science One), design and deploy interventions on online platforms. To that end, I am actively collaborating with teams at Meta, Google, Mozilla supporting me via grants and data access, and working in the Indian context with newsrooms, Deutsche Welle and Jagran Media, jointly with a reach of over 200M readers on field experiments deploying civic technologies. My team and I also launched an MIT-incubated tech venture deploying a multilingual health-literacy product in India that has received awards from UNICEF, MIT, Goethe Institut, and is live with 300+ real users in Jalgaon, Maharashtra, and Dhaka, Bangladesh, allowing us to test IRB-approved digital healthcare interventions with local nonprofits supporting us. We are designing systems to evaluate offline behavioral changes induced by the AI-informed interventions, scaling to 5000 users by 2025.

In the last year at the Integrity Institute, an online safety and policy think tank, I have contributed significantly to their work on platform transparency recommendations, elections integrity best practices, Online Safety Act consultations by Ofcom, and the Digital Services Act consultations by the European Commission (ECAT), resulting in an invitation to join their Community Advisory Board. This also informed SimPPL's work drawing on historical transparency policies and their outcomes in healthcare, finance, and aviation, to provide evidence that could guide regulators looking into AI and online safety [17]. I spoke at panels on partnerships, AI auditing, and disinformation hosted by the Finnish Embassy, UNESCO, TrustCon, Humane Intelligence, where I brought in my experience deploying AI-based tools in underserved regions in India, Bangladesh, Mongolia to explain why existing auditing and policies are inadequate for consumer safety. Through my research I saw uninformed mothers going back into the fields to work in a few hours of childbirth, adolescent girls suffering as a result of menstrual health misinformation, and fact-checkers spending hours attempting to manually identify misinformation on Facebook one link after another. These experiences influence my view of the problems we need to ensure technology and policy can improve the status quo for. I intend to drive substantive societal change through my research in the social sciences and human-centered computing.

[7] To solve this problem, there are a number of 'national observatories' being created as sources of rich mobile and web activities from consenting users via browser extensions–with whom we are starting to collaborate

[8] An important pragmatic consideration for my long-term research may be the influence of changes in U.S. political leadership that affects funds available for U.S.-focused misinformation research. I hedge this with both, my focus on digital trust, and global partnerships.

References

[1] A. Shah, P. Banerjee, J. Kuang, A. Nichols, N. Mazar, M. van Alstyne, S. Mehta. **'Market Design Interventions for Safer Agentic AI'**. *Working Paper in Preparation for ISR; Oral Talk, Yale AI, ML, and Business Analytics Conference, 2024.*

[2] H. Ranka, D. Shah, R. Sannikov, E. Brichetto, L. Ng, S. Mehta. **'Bridging Nodes and Narrative Flows: A Graph-Theoretic Analysis of Telegram's Disinformation Ecosystem'**. *Working Paper, in preparation for ICWSM; Poster at Stanford T&S Research Conference, 2024.*

[3] S.W. Baksh, D. Mungra, V. Pariawala, R. Jain, A. Surve, A. Das, M. Biswas, F. Afroz, S.M. Eron, S. Mahmood, S. Mehta, S.R. Diya. **'From homophobia to assault: The gendered landscape of Bangladesh's political disinformation'**. *Report (Tech Global Institute), 2024; Meta issued takedown.*

[4] S. Mehta, M. van Alstyne. **'Improving Economic Welfare in Reputation-based Marketplaces with Truth Warrants'**. *Working Paper. Oral Talk, Workshop on Information Systems and Economics, 2024.*

[5] S. Mehta, Z. Yang, C. Fan, D. Liu, N. Mazar, M. van Alstyne. **'Reducing Misleading Claims in Digital Marketplaces using Truth Warrants'**. *Working Paper in preparation for CHI (Late Breaking), 2025.*

[6] G. Malpani, S. Mehta. **'The Effect of Influencers on the Independent re-sharing of News in an Alt-Right Ecosystem'**. *Working Paper in preparation for PNAS Nexus, 2024.*

[7] S. Kalichman, L. Eaton, V. Earnshaw, and N. Brousseau. Faster than warp speed: early attention to COVD-19 by anti-vaccine groups on Facebook. *Journal of Public Health, 2022*,

[8] V. Dalal, A.V. Singh, C. Schroeder de Witt, S. Mehta. **'Using Large Language Model Debates to Predict X's Community Notes Visibility'**. *Working Paper, 2024.*

[9] R. Jain, S. Mehta, F. Barez, P. Torr. **'Detecting Emergent Social Media Misinformation in the Era of LLMs'**. *Working Paper in preparation for ACL, 2024.*

[10] S. Mehta, J. Bisbee, Z. Sanderson, R. Bonneau, J. Tucker, and J. Nagler. **'Identifying the Causal Effects of Twitter's Interventions on Trump's Tweets'**. *Working Paper in preparation for PNAS; Oral Talk, Stanford Trust and Safety Conference; Invited Talk, Twitter 2022-24.*

[11] S. Mehta, K. Aslett, Z. Sanderson, R. Bonneau, J. Tucker, and J. Nagler. **'Platform Outages Increase the Sharing of High-quality News on Reddit'**. *Working Paper, 2023.*

[12] S. Mehta, S. Sarkhel, X. Chen, S. Mitra, V. Swaminathan, R. Rossi, A. Aminian, H. Guo, and K. Garg. **'Open-Domain Trending Hashtag Recommendation for Videos'**. *In IEEE International Symposium on Multimedia (ISM), 2021.*

[13] S. Mehta, A.G. Baydin, R. Bonneau, J. Nagler, and P. Torr. **'Estimating the Impact of Coordinated Inauthentic Behavior on Content Recommendations in Social Networks'**. *Oral Talk, Workshop AI for Agent-Based Modelling, ICML 2022.*

[14] A. Nichols, N. Mažar, T. Parker, S. Mehta, G. Pennycook, D. Rand, and M. Van Alstyne. 'Certifiably True: The Impact of Self-Certification on Misinformation'. *Working Paper, In Preparation for PNAS*

[15] C. Shah, R. Konka, G. Malpani, S. Mehta, and L. Ng. **'Can Social Media Platforms Transcend Political Labels? An Analysis of Neutral Conservations on Truth Social'**. *Oral Talk, DARE Workshop, AAAI Intl. Conference on Web and Social Media (ICWSM), 2024.*

[16] M. van Alstyne. 'Free Speech, Platforms & The Fake News Problem', 2022.

[17] C. Vergara, R. Jain, and S. Mehta. **'A History of Transparency Reform to Shape Platform Regulation'**. *Intl. Conference on Digital Governance (dg.O), 2024.*

[18] L. Tay, S. Lewandowsky, M. Hurlstone, T. Kurz, U. Ecker. 'A focus shift in the evaluation of misinformation interventions'. Harvard Kennedy School Misinformation Review, 2022.

[19] L. Liu, and B. Weingast. ‘Taobao, federalism, and the emergence of law, Chinese style.‘; Minn. L. Rev. 102 (2017): 1563.