MIT MANAGEMENT SLOAN SCHOOL

Rising Scholars Conference
Information Systems Student Research Presentations

Jonathan Gomez Martinez	**Emory University**
jonathan.gomez.martinez@emory.edu	Goizueta Business School

Jonathan Gomez Martinez is a PhD candidate in Information Systems and Operations Management at the Goizueta Business School. Informed by his background as a Mexican immigrant and first-generation college student, Jonathan's research highlights the unintended consequences of technology and technology policy. His ongoing projects evaluate the role of AI, privacy policy, and digital platforms on censoring minority voices and complicating the operations of small and midsized businesses. To learn more about Jonathan, visit www.jgomezm.com.

Abstract:

Platform Policy Changes: Impact of Auto-Moderation on Minority Community Rights

User-generated content on social media platforms has always been moderated as advertisers on these platforms require interactions to be safe, non-abusive, and generally in compliance with regulations such as those dealing with intellectual property rights. Even if assisted by algorithms to filter content, human reviewers had always made the final call, until recently where unprecedented volume and other factors have forced platforms to rely fully on automated, Artificial Intelligence based (AI-based) content moderation. Cognizant of unintended consequences of technology usage, our research exploits a natural experiment wherein Twitter had resorted to auto-moderation in 2020. Our investigation reveals the dramatic impact of such technologies, often context-blind, on the interactions of a minority group of users such as the LGBTQ+ community. Through a rigorous empirical approach, our findings show that interactions within this community reflect a heavily censored language after auto-moderation deployment by Twitter. In the absence of any explicitly LGBTQ+ related policy changes on Twitter, our work underscores the inadvertent harm that ensues when context-less AI technologies are adopted.

Farnam Mohebi
farnam.mohebi@gmail.com

Univeristy of California, Berkeley
Haas School of Business

I am currently a management PhD student at the Haas School of Business and a data science fellow at the Dlab, UC Berkeley, having previously completed my MD-MPH. I focus on the intersection of healthcare and management, driven by a deep interest in understanding the multi-faceted role of physicians in the AI world. I am interested in physicians' perception and experience with clinical AI and physician-scientists' narratives of it. Additionally, I study the impact of management practices on physicians. My work is guided by my background in healthcare and a commitment to improving organizational practices within the field.

Abstract:

Assessing the Multifaceted Role of Physicians in the AI Landscape.

1 Research Question

The central research question of this study is multi-faceted, exploring how physicians navigate the rapidly evolving landscape of Artificial Intelligence (AI) in healthcare. Specifically, the question aims to unpack the complexities of physicians' diverse roles as developers, adopters, evaluators, and managers of AI technologies in medical settings.

As Developers: Why are Physicians Becoming Developers? How do physicians influence the trajectory of medical AI science development? Beyond their influence on the trajectory of AI, how does the professional standing of physicians contribute to their credibility as developers? Do non-financial incentives, such as academic recognition or potential for societal impact, also play a role? What issues of legitimacy arise when physicians act as developers? Are there elements of elitism and prestige that attract physicians to the field of AI development? How do these factors interact with other motivations and constraints?

As Adopters: How do elements like professionalism and hierarchy affect physician adoption rates at various levels?How do age, gender, and other demographic characteristics of physicians influence their willingness to adopt AI technolo-gies? Does a younger generation of physicians, for example, show less openness to incorporating AI into their practices compared to their older counterparts? How do the level of training and the years of clinical experience impact a physician's propensity to adopt AI? Do physicians with more advanced training or specialization show different patterns of adoption?What role does exposure to AI in medical education play in facilitating or hindering adoption? How is the legitimacy of the technology assessed before adoption? Are there non-financial incentives that significantly impact adoption, such as the prospect of improved patient outcomes, peer recognition, or professional development opportunities?

As Evaluators: In what ways do physicians' professionalism ensure a more rigorous and ethically sound evaluation of AI tools? Does the commitment of physicians to professional ethics and existing medical practices make them more resistant to adopting innovative AI technologies that challenge traditional healthcare paradigms? Are senior physicians or those higher up in the medical hierarchy more likely to maintain the status quo, thereby hindering the adoption of transformative AI technologies?

As Managers: How does a physician’s role as a manager facilitate the integration of AI technology into clinical settings, particularly in terms of operational efficiency and patient care? In what ways does the managerial role of physicians contribute to fostering an organizational culture that is more receptive to AI innovations? How does a physician-manager’s clinical background influence the prioritization of AI projects that have the most direct impact on patient care?Does a physician’s managerial role lead to conflicts of interest when deciding on AI projects, perhaps prioritizing those that align with their own clinical specializations or authority over others that may benefit the healthcare system more broadly? How might the dual responsibilities of physician managers contribute to potential burnout, thereby affecting their capacity to evaluate and implement AI technologies effectively?

2 Methods

My research adopts a full-cycle approach. In the initial qualitative stage, I will employ ethnography and content analysis to delve into various social media, press releases, scholarly publications, and other publicly available data that provide insights into physicians’ perspectives on AI. The focus will be on their roles as developers, adopters, evaluators, ethicists, and managers.

The qualitative insights will then inform the design of lab or field experiments, administrative data analysis, and surveys. These will focus on physicians’ decision-making patterns, adoption rates, ethical considerations, and managerial choices in the context of AI integration into healthcare.

3 Implications

Understanding the economic and social dynamics that influence AI adoption is crucial for policymakers and industry stakeholders. This research could contribute valuable insights into how to navigate the conflicting interests between organizational efficiencies and end-user acceptance. It also opens up discussions on the economic implications of technology adoption in critical sectors like healthcare.

Carolina Reis **Virginia Tech**
creis2@vt.edu **Pamplin College of Business**

Carolina Reis is a fourth-year Information Systems PhD student at Virginia Tech. Broadly, her research focuses on the hybrid human-machine behavior. In particular, her research investigates: (1) how the introduction of AI systems into social and organizational ecosystems alters human beliefs and behaviors, and (2) how people themselves also shape AI systems through the training of these systems using active human input.

Abstract:

Outsourcing Morality: The Hidden Path to Machine Ethics

Authors: Carolina Reis, Virginia Tech; Nicholas Brown, Indiana University

Artificial intelligence (AI) technologies rely on large language models (LLMs) trained on easily accessible information online, such as content from the “darkest recesses of the internet”

(Perrigo, 2023). To remove toxicity (e.g., sexism, racism, xenophobia, hate speech, calls for violence) from the training sets, companies hire human agents to perform content moderation—the process of reviewing and monitoring digital toxicity. Currently, a large share of this content moderation process is outsourced to companies and workers in developing nations, where the work is often unregulated, leaving content moderators to label and annotate toxic content on their own. These subjective labels then serve as the ground truth for LLM technologies (Perrigo, 2023). However, the appropriateness of language use varies among cultures, contexts, and people, and what is morally acceptable depends on where the person lives. A paradox thus ensues: AI technologies are used worldwide, especially in technologically advanced countries, but their ingrained morality is determined in foreign countries that do not necessarily hold similar moral values.

In this research, we intend to investigate the differences in cultural perspectives that influence content moderation and whether these differences perpetuate harmful AI biases. To begin, we conducted a pilot study, using the leaked Facebook documents on hate speech content moderation (“Hate Speech and Anti-Migrant Posts: Facebook’s Rules,” 2017), to assess whether individuals from different countries vary in their moral predisposition. Preliminary results confirm this hypothesis, and show that (1) a model powered by content moderated by individuals from the United States (n = 391) would have a significantly higher accuracy rate, precision rate, recall rate, and F1 score than a model powered by content moderated by individuals from India (n = 286), and (2) a model powered by content moderated by individuals from India would have a significantly higher accuracy rate, precision rate, recall rate, and F1 score than a model powered by content moderated by individuals from Brazil (n = 32), when the Facebook guidelines are used as ground truth. These results indicate cross-cultural ethical variation and raise potential concerns with current machine ethics practices.

In our forthcoming studies, we will employ a mixed-method design. In Study 1, we will interview content moderators located in different countries to grasp the practices adopted in the content moderation process. In Study 2, we will launch an online experimental platform (similar to Awad et al., 2018) where we will explore different content moderation scenarios and collect data from

people in multiple countries to assess cultural differences in ethical tendencies. In Study 3, we will develop algorithmic models powered by the labeled data from different cultures and show the differences in algorithmic output and performance. Ultimately, our aim is to help advance potential solutions for the problem of universal machine ethics.

Kai-Cheng Yang
yang3kc@gmail.com

Northeastern University
Network Science Institute

Kai-Cheng Yang is a postdoctoral researcher in the Lazer Lab at Northeastern University's Network Science Institute. He obtained his Ph.D. in Informatics from the Luddy School of Informatics, Computing, and Engineering at Indiana University Bloomington. He is interested in computational social science. His research aims to uncover how technologies like generative AI are used for deceptive and disruptive purposes, study how humans react to these abuses, and develop countermeasures. Specifically, he focuses on bad actors like malicious social bots and misinformation on social media. He built popular tools, such as Botometer, that have served tens of thousands of users. He also acted as the social bot expert in the trial of Twitter vs. Elon Musk. His work has been covered by CNN, BBC, The New York Times, and many other popular news outlets.

Abstract:

Large language models and cyber social threats: Good, bad, and ugly

Large language models (LLMs) may profoundly impact our information ecosystem. On the one hand, they exhibit impressive capabilities in generating realistic text across diverse subjects and show great potential in many applications. On the other hand, concerns have been raised that they could be utilized to produce fake content with deceptive intentions. In this talk, I will present three studies to demonstrate how LLMs can be abused by bad actors and leveraged by users for self-protection. In the first study, I will introduce a Twitter botnet that appears to employ ChatGPT to generate human-like content. These accounts form a dense cluster of fake personas that exhibit similar behaviors, including posting machine-generated content and stolen images, and engage with each other through replies and retweets. ChatGPT-generated content promotes suspicious websites and spreads harmful comments. While the accounts in the botnet can be detected through their coordination patterns, current state-of-the-art LLM content classifiers fail to discriminate between them and human accounts in the wild. In the other two studies, I will talk about using LLMs to counter the spread of misinformation. Through extensive experiments, I find that ChatGPT, a prominent LLM, can evaluate the credibility of news outlets. This suggests that LLMs could be an affordable reference for credibility ratings in fact-checking applications. Then, I further test the feasibility of using ChatGPT as a fact-checking tool in a human-subject experiment. Although ChatGPT performs reasonably well in debunking false headlines, it does not significantly affect participants' ability to discern headline accuracy or share accurate news. In certain cases, ChatGPT might even be harmful. The findings underscore the importance of accounting for human factors when incorporating AI models into our information ecosystem.