MANOEL HORTA RIBEIRO RESEARCH STATEMENT

My research aims to understand and improve **moderation**, **recommendation**, and **monetization** practices in **online platforms** by using and developing methods in **computational social science**, **causal inference**, **network analysis**, **natural language processing**, and **machine learning**.

To improve online spaces, we must ask *what if?* What if we 'deplatform' a far-right website? Or what if we delete comments inciting violence? **These are inherently causal questions that descriptive work cannot answer alone.** Thus, to achieve my research goal, I collect observational or experimental data online and then analyze and model them, **generating causal insights that can inform online platforms' design and policy**. This undertaking is well aligned with my computational social science background and my previous experiences working with Meta, Microsoft, Reddit, and Prolific. In some cases, drawing causal conclusions requires deep technical knowledge of how these platforms work, e.g., understanding how machine learning algorithms are used in content moderation; in others, it requires expertise in causal inference, e.g., finding quirks in observational data that help identify causal effects.

Throughout my doctorate, **I have independently created and pursued my own research agenda**, spearheading multiple collaborations with companies, other researchers, and non-profits to explore how content curation practices can enhance online spaces. For example, I collaborated with researchers from Meta and Reddit to study the effect of content moderation (WWW'23; ICWSM'22); with peers from Cornell Tech to study monetization strategies on YouTube (CSCW'22); and, currently, with Tournesol, a Swiss non-profit, to investigate whether browser extensions can help people improve their social media diets. This wide net of collaborations led to award-winning (CSCW'21), influential papers, a Facebook Fellowship, a Forbes 30 under 30 award, and extensive media coverage, with my research appearing in over 70 news pieces from outlets like Bloomberg, DER SPIEGEL, and NBC News.

My research has informed systems that process millions of posts and videos every day. My work on online radicalization on YouTube (the 3rd most influential paper in CS in 2019, per Altmetrics) spawned a rich literature on algorithmic effects and led to multiple invited talks at Google (FAccT'20). My work showcasing the effectiveness of post approvals (ICWSM'22), a feature where moderators require posts in Facebook groups to be pre-approved, led to **expanding the feature** to comments and prioritizing additional features that decreased the effort to approve posts. My research on content removals (WWW'23) has **broadened the success metrics** associated with content moderation within Facebook—leading them to consider second-order effects like subsequent rule-breaking behavior when moderating content. Beyond social media, my work with Prolific has **informed the crowd work platform** and researchers at large about the use of large language models by crowd workers, as well as how to mitigate said use (arXiv:A).

Online platforms like Facebook, YouTube, and Reddit have changed the fabric of society, from facilitating activism around climate change to undermining public health policy during the COVID-19 pandemic. Positive or negative, the effects of these platforms on humankind are mediated by how they moderate, recommend, and monetize content through a mix of policies, human labor, and algorithms. However, compared to other areas of great public interest, like economic policy or public health, **how we design and curate content on online platforms is still ill-informed by rigorous academic research**. In what follows, I describe two recent research projects that outline my vision for research that bridges this gap: informing content curation and improving the Web.

DOES DEPLATFORMING WORK? (PNASNex‘23)

To curb false information, hateful speech, and conspiracy theories, platforms have banned (or ‘deplatformed’) influencers, communities, and even entire websites. **Yet, the impact of deplatforming on the Web remains unclear**. Sanctioned individuals migrate to more permissive and less public-facing websites where their engagement is harder to track. Further, deplatforming often occurs in response to an event—and we must isolate what is caused by deplatforming *vs.* the associated event.

◇ **Case study: Parler.** I addressed the challenges of studying deplatforming in the context of a high-profile deplatforming event: the suspension of the popular US social networking service Parler from Amazon’s Web hosting services on January 11, 2021, following the US Capitol attack. We found that **the deplatforming of Parler drove users to other fringe social media**, like Gab, Rumble, and 4chan, such that, in total, **there was no overall decrease in activity** on fringe platforms. Given that the isolated deplatforming of a major fringe platform was ineffective, we recommended stakeholders consider alternative courses of action, like promoting simultaneous action against multiple fringe social media platforms or acting proactively rather than reactively during periods of political unrest.

Obtaining this crisp finding with clear policy implications was **only possible using the right data and methods**. We used two online panels from The Nielsen Company capturing passive consumption on desktop (n=6,677) and mobile (n=36,028) devices. This data allowed the **tracking of passive engagement** with fringe content across **various websites and apps**. To disentangle changes in fringe consumption caused by the US Capitol attack, we used a difference-in-differences approach; see Fig. 1. In essence, we **avoid the confounding effect** of the attack by comparing two populations similarly impacted by it: users active on Parler and users active on other fringe social media.

The effectiveness of **interventions like deplatforming must be assessed** to understand whether they can improve the Web. **My research provides stakeholders with this knowledge**, rigorously mapping interventions to expected outcomes and **guiding policy away from guesswork**.

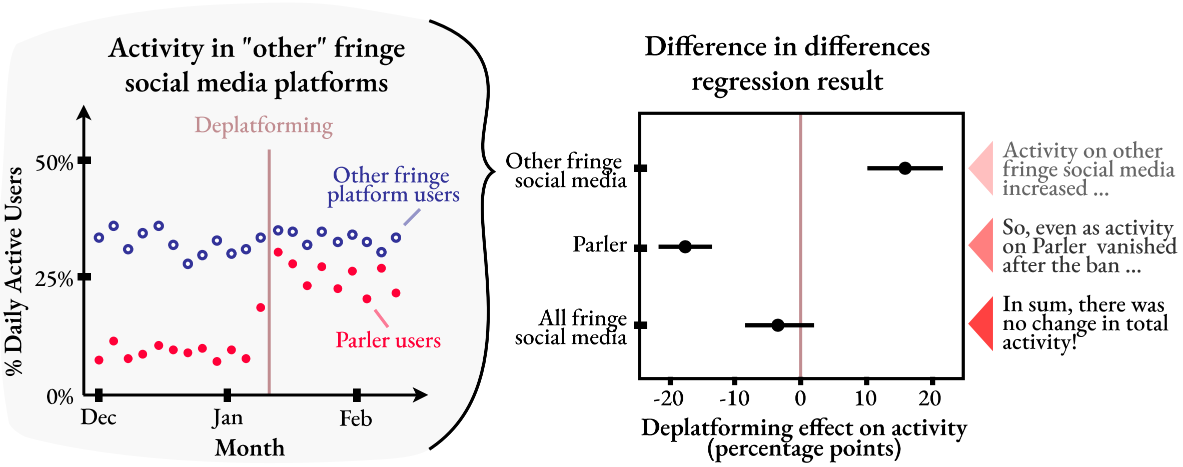

Fig. 1: To estimate the causal effect of deplatforming on user activity, I compared users who consumed mostly Parler (in red on the left) with users who consumed mostly other fringe social media (e.g., Rumble, Gab; in purple), matched for demographics and overall fringe consumption pre-deplatforming. Under the parallel trends hypothesis, i.e., that consumption differences would have remained constant in the absence of the deplatforming, we can estimate the causal effect of deplatforming on user activity for Parler users (estimates are shown on the right). **We found that activity on other fringe social media increased so that, even with the banning of Parler, there was no change in total activity.**

DOES CONTENT MODERATION IMPROVE USER BEHAVIOR? (WWW‘23)

Most content moderation efforts concern **micro-level decisions** about millions of posts, comments, images, and videos, decisions that, together, **fundamentally shape online platforms**. Since the inception of online platforms, content curation practices like *when to delete a comment* have been studied mainly within companies like Meta and Google. Part of my research agenda is to **study content curation in academia**, where we have the **openness and rigor** necessary to develop sound research and can meaningfully inform *all* stakeholders, from platforms themselves to legislators.

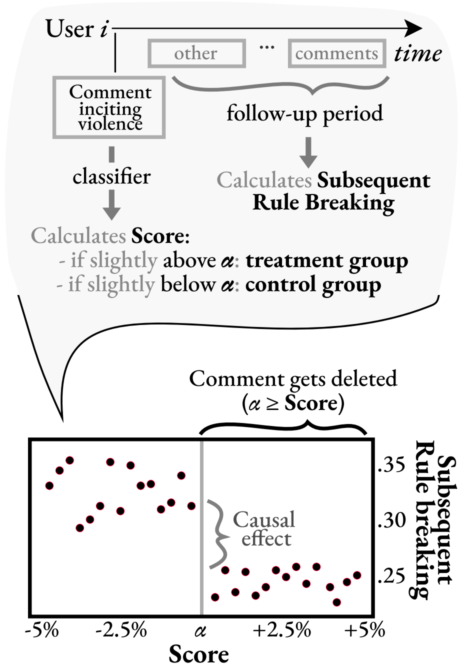

Fig. 2: For each user, we consider the first comment whose score was close to the threshold α, at which a moderation intervention is enacted and outcomes calculated in the weeks after; here, the number of subsequent rules broken (see top). Aggregating over users (see bottom), we disentangle the causal effect of the intervention by measuring the discontinuity in subsequent rule-breaking (y-axis) around the threshold α considering the scores of users' first comments (x-axis). **Here, the deletion of a comment led to a decrease in the number of subsequent rule breaking.**

◇ **Case study: Facebook comments.** On Facebook, millions of posted comments go through machine learning classifiers that trigger moderation decisions. These **classifiers prevent harm** by deleting rule-breaking content—**but do they make users behave better in the long run?** Ideally, we could run a large-scale randomized experiment to answer this question: whenever a comment is classified as harmful, we could flip a coin to decide whether to remove it. Yet, such **an experiment would be troublesome ethically**, as platforms would knowingly abstain from preventing harm and experiment without consent.

Comparing users with vs. without deleted comments is also unlikely to yield answers, as differences observed might stem from differences between users rather than the effect of moderation. **I overcame these challenges in a large study** (n=412M) with a quasi-experimental research design (WWW‘23). Moderation decisions are taken when scores given by classifiers exceed an arbitrarily chosen threshold α. Around this threshold, comments are very similar, but those with scores slightly higher than α are moderated, whereas those with scores slightly lower are not. Thus, **I could simulate a randomized control trial around the threshold**, isolating confounders and obtaining a sharp causal estimate; see Fig. 2. Using this ‘regression discontinuity,’ I found that **deleting rule-breaking comments led to a transient decrease in engagement and a persistent decrease in rule-breaking behavior**.

Automated content moderation often struggles to incorporate the full cultural and situational context of online content, being, at times, too blunt. Yet, in a nutshell, **my research indicates that it broadly achieves its goal**, a piece of information critical for stakeholders inside and outside of Meta to consider when weighing the pros and cons of such systems.

My work on deplatforming and content moderation illustrates my way of doing research. I like to draw **causal conclusions** from **massive data;** often by finding **natural experiments**. I enjoy tackling **real-world issues** and providing **actionable insights**, and I often accomplish this by working closely with **industry partners**.

RESEARCH AGENDA

My research aim is to provide actionable insights to improve online platforms, guiding policy, design, and practice. To do so, it is crucial to **understand how new technologies and practices will shape the Web** and **how they can expand what can be studied**.

◇ How to improve algorithmic auditing? Research should inform stakeholders about the impact of algorithms on platforms like YouTube and TikTok. Yet, **evaluating the causal effect of recommender systems is challenging**, as the counterfactual world where they do not exist is ill-defined—what would TikTok be without a recommender system? In recent work under R&R at PNAS alongside peers from The University of Pennsylvania, I created 'counterfactual bots,' automated agents that simulate user behavior up to a certain point, at which an intervention may change their behavior (arXiv:B). Comparing the recommendations from bots that at some point started blindly following recommendations with those from bots that did not, we **disentangled the causal effect of the YouTube recommender system**—which we found drives people *away* from extreme content.

In the future, I want to conduct further algorithmic audits considering this complex interplay between human preferences and algorithms. I am particularly interested in using large language models (LLMs) to help in these audits. **Given that LLMs are remarkable at simulating human behavior, could we use them to conduct algorithmic audits?** For example, could we study the impact of social media on teens' mental health? The scarcity of representative online traces limits the credibility of algorithmic audits, e.g., the 'counterfactual bots' methodology only works because we have (expensive) online traces from The Nielsen Company. With LLMs, I hope to democratize and enhance algorithmic auditing.

◇ How will LLMs impact online platforms? Beyond their utility in computational social science, **large language models will shape the Web in the coming years**. For example, in recent work, **I showed that LLMs are widely used on crowd work platforms** and that targeted mitigation strategies can reduce, but not eliminate, their usage (arXiv:A). This is concerning as **LLM use can threaten the validity of research conducted on crowd work platforms** like Prolific and Amazon Mechanical Turk, as researchers usually care about human (rather than model) behavior or preferences.

Understanding how large language models will impact social media is even more urgent. These models can browse the Web, engage in conversation in human-like fashion, and interact with multimodal content; thus, they **are bound to create new challenges to online platforms**. Small radical communities already capable of shaping media narratives and harassing users will use these models to astroturf and artificially generate content capable of misleading and harming others.

In future work, **I want to measure and mitigate potential harms coming from LLMs**. I am particularly interested in the impact of LLM-generated propaganda and sockpuppetry. Using **experiments**, I want to measure LLMs' persuasiveness (compared to other humans) in a 'short debate' setting mimicking social media discussions. But beyond focusing on one particular harm, I want to lead a broad initiative to track LLM 'aggregate' prevalence in online platforms (a task simpler than detecting individual cases). This 'LLM observatory' would **enable observational studies measuring the impact of LLMs** and of various interventions as the technology and associated policies co-evolve.

Online platforms are a mix of humans and machines. My work shows that research by **computational social scientists** like myself can inform their design and policy. As new technologies like LLMs grow in capability and popularity, I believe that **bridging the gap between social and computational research** will be needed to reap the benefits and prevent the harms from online platforms.

References

[arXiv:A] V. Veselovsky*, M. Horta Ribeiro*, P. Cozzolino, A. Gordon, D. Rothschild, and R. West. Prevalence and prevention of large language model use in crowd work. *arXiv preprint arXiv:2310.15683*, 2023.

[arXiv:B] H. Hosseinmardi, A. Ghasemian, M. Rivera-Lanas, M. Horta Ribeiro, R. West, and D. J. Watts. Causally estimating the effect of youtube's recommender system using counterfactual bots. *arXiv preprint arXiv:2308.10398*, 2023.

[CSCW'21] M. Horta Ribeiro, S. Jhaver, S. Zannettou, J. Blackburn, G. Stringhini, E. De Cristofaro, and R. West. Do platform migrations compromise content moderation? Evidence from r/The_Donald and r/incels. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), 2021.

[CSCW'22] Y. Hua*, M. Horta Ribeiro*, T. Ristenpart, R. West, and M. Naaman. Characterizing alternative monetization strategies on youtube. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–30, 2022.

[FAccT'20] M. Horta Ribeiro, R. Ottoni, R. West, V. A. Almeida, and W. Meira Jr. Auditing radicalization pathways on youtube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 131–141, 2020.

[ICWSM'22] M. Horta Ribeiro, J. Cheng, and R. West. Post approvals in online communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 2022.

[PNASNex'23] M. Horta Ribeiro, H. Hosseinmardi, R. West, and D. J. Watts. Deplatforming did not decrease parler users' activity on fringe social media. *PNAS Nexus*, 2(3), 2023.

[WWW'23] M. Horta Ribeiro, J. Cheng, and R. West. Automated content moderation increases adherence to community guidelines. In *Proceedings of the ACM Web Conference*, 2023.