JEREMY FOOTE—CAREER: DEVELOPING COMPUTATIONAL APPROACHES TO UNDERSTAND BELIEF CHANGE AND RADICALIZATION IN ONLINE COMMUNITIES

Project Summary

Online spaces are vital resources, serving as sources of information, technical support, and social support for millions of people. However, online spaces are also plagued by a number of problems, including belief radicalization and toxicity. These problems contribute to a lack of safety online and increasingly threaten public safety and political stability offline. While earlier research has focused on algorithmic curation or affordances of social media as explanatory factors, work from psychology on persuasion and radicalization, as well as recent work on online radicalization, suggests the importance of centering *online communities* as drivers of belief and behavior change. Online communities—online spaces organized around a shared topic—can provide a shared sense of identity and purpose to their members: ideal conditions for persuasion. Existing research on this topic has typically focused on belief or behavior change in a single community. However, online community platforms are composed of many overlapping, interdependent communities, which users rapidly move between. These affordances almost certainly influence how identity formation, exposure, and radicalization occur in these spaces. Understanding these phenomena requires applying theories and data that encompass entire online community ecosystems.

The proposed work will provide new approaches to answer questions about the role that online communities have in changing people's beliefs and behavior, through a multi-pronged approach. At the individual user level, it will use large-scale, longitudinal data to identify platform users who have changed beliefs through identifying 1) explicit expressions of belief and 2) activity in belief-oriented communities (e.g., 'r/conspiracy'). We will use a similar approach to identify users who change behavior, such as beginning to use (or discontinuing use of) slurs or highly toxic language. After using qualitative methods to validate this approach, we will compare the history of these users with others who look similar but did not change beliefs or behavior in order to identify indicators and potential causes of belief and behavior change. At the community level, the project will create a participation network, based on identifying the temporal order in which people join communities, thus identifying communities that might act as "pathways" toward or away from radical or otherwise problematic communities.

Descriptive findings from these two projects, combined with other theories and empirical evidence in this space, will inform the development of projects using natural experiments and field experiments to test the role of exposure and joining processes in promoting belief and behavior change. For example, we will be able to identify communities which are themselves not problematic but which are on pathways toward problematic communities. Working with moderators of these communities, we will develop interventions aimed to reduce exposure to the problematic communities and test whether these interventions change the outcomes of community members. At an individual level, we will be able to identify users who are at-risk of radicalization (or deradicalization) and test interventions such as recommending positive communities.

The **intellectual merit** of the proposed work includes deeper theoretical explanations for the role that online communities play in belief formation and belief change, as well as causal evidence for some of these explanations. In addition, the work will develop methodological advances in longitudinal modeling of belief formation and community migration, as well as developing approaches for applying natural language processing and causal inference in this setting.

The **broader impacts** of the work include developing both theory and methods for identifying when online communities are harmful, and for whom. The findings will also provide initial evidence of strategies which may be effective in deterring dangerous beliefs or behavior; knowledge which will be directly applicable for community managers and policy makers as they seek strategies to build safer, more productive online spaces.