

OWEN BROWN

DETERRING HATE SPEECH ONLINE: FUEL OR FAIL? EMPIRICAL EVIDENCE OF COORDINATED COUNTER SPEECH

Social media have become ubiquitous, and nearly every segment of the global population can benefit from these platforms. However, social media (SM) also have a substantial side effect: they significantly increase individuals' exposure to hate speech. While policymakers have sought to address this problem, there is no consensus on which moderation policy is most adequate. Moderation policies such as content curation or speech deletion raise concerns because they may jeopardize freedom of expression. In contrast, counterspeech—communication intended to counteract the potential harm caused by other speech—has recently attracted growing attention. Although prior research shows that counterspeech can be effective when it targets individuals, the broader effectiveness of this moderation strategy remains underexplored. In particular, existing studies have not examined the causal impact of counterspeech on hateful online public debate. This paper aims to fill this gap. Exploiting a quasi-natural experiment on Facebook over a 15-month period, I investigate whether and to what extent a coordinated counterspeech intervention reduces the level of hatefulness in published messages.