12-12-2022

Can human complement an erring algorithm? Assessing Human-Bot Hybrid Designs for Managing Online Discussions

Xinyu Fu
University of Pittsburgh, xinyu.fu@pitt.edu

TREO
Technology, Research, Education, Opinion

Can human complement an erring algorithm?

Assessing Human-Bot Hybrid Designs for Managing Online Discussions

Xinyu Fu, University of Pittsburgh, Xinyu.fu@pitt.edu

Online communities often face the challenge of trying to remove inappropriate content posted by visitors, yet simultaneously attempting to promote helpful user participation. As communities have grown exponentially in membership over the last two decades and have spread to encompass every time zone on the globe, human moderation has become too slow, expensive, and difficult to manage. Digital platforms have responded by adopting learning-based systems (bots) to assist with content moderation. Compared with traditional rule-based systems, these bots have the potential to continuously improve their performance with end user participation, but they are also less transparent, and their errors are less predictable. We examine how end users utilize bot advice in the online communities change in the presence of such a highly accurate but still imperfect bot. Specifically, we tested under what conditions users could effectively detect the bots' errors, and override the bots' assessment where the bot makes mistakes. Drawing insights from Error Management Theory and automation literature, we present *error anticipation* as a key condition for improving the detection rate for algorithmic errors, which is mediated by *complacency potential* (i.e., the tendency to alleviate workload on machines without monitoring for potential algorithmic errors). To trigger users' error anticipation, we employed an explanation technique in which they were asked to explain how the bot reached specific conclusions on exemplar comments before they were shown on which exemplar comments the bot made errors. In three waves of experiments, we simulated an online news discussion forum where the content moderation task was crowdsourced to users assisted by a bot. We found that, on average, users aided by the bot achieved higher moderation decision quality than users unaided by the bot. Users who provide an explanation about bots' assessments on exemplar comments, however, were able to better detect the bot's errors and achieve higher performance than users who are (1) generally informed that the bot is imperfect and (2) directly told where the bot is erring in the given exemplar comments. Furthermore, complacency potential partially mediates the effects of error anticipation on error detection. Overall, these results suggest that encouraging error anticipation via explanation could help users become more vigilant about the actions of imperfect bots. Based on these results, we discuss the theoretical and practical implications of deploying human-bot hybrid designs in digital platforms.