American Psychological Association

Technology, Mind & Society
2021 Conference Proceedings

© 2021 The Author(s)
ISBN:

http://dx.doi.org/10.1037/xxxxxxx

AI is Blamed More Than Humans When it Falters, But Less so When it Discriminates*

Tripat Gill[1] , Shirley Chen[1], and Sakshi Korde[1]
[1]Wilfrid Laurier University

While people are reluctant to use AI (artificial intelligence)-based decision agents, little research has examined blame assignment when such agents falter. The *type* of fault – performance error versus discrimination by the agent – has also not been examined. These issues were investigated in two experimental studies. In study 1 an AI versus human medical screening agent makes an accurate versus an inaccurate diagnosis (positive vs. negative outcome, respectively). There was no difference in credit assigned to the agents for positive outcomes. But the AI agent was assigned more blame than the human for negative outcomes. Study 2 replicated this effect in a job recruitment scenario and also examined two more conditions that specified the type of fault: the agent made a *performance* error or it *discriminated* based on the gender of the applicants. While both agents were equally credited for positive outcomes (job application accepted), AI was blamed more than the human recruiting agent for negative outcomes (job application rejected). However, this difference vanished in the performance error condition and was significantly reversed in the discrimination condition. Specifically, the AI agent was blamed less than the human recruiter if it discriminated. This reversal in blame assignment was mediated through higher moral outrage towards human versus AI agents that discriminate. Thus, biased AI agents may continue to prevail as they garner lower moral outrage and are assigned lower blame for their discriminatory behavior.

Keywords: artificial intelligence, blame, moral outrage, performance errors, discrimination

This article was published [to be completed by publisher].
https://orcid.org/0000-0001-9839-4113
This research was supported in part by Social Science and Humanities Research Council of Canada (grant # 155826). The authors have no conflict of interest to disclose.
Correspondence concerning this article should be addressed to Tripat Gill, Wilfrid Laurier University, Waterloo, ON, N2L3C5, Canada. Email: tgill@wlu.ca