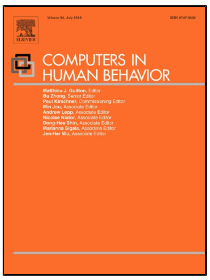
Accepted Manuscript



Attributions of Morality and Mind to Artificial Intelligence after Real-World Moral Violations

Daniel B. Shank, Alyssa DeSanti

PII:	S0747-5632(18)30240-1
DOI:	10.1016/j.chb.2018.05.014
Reference:	CHB 5522
To appear in:	*Computers in Human Behavior*
Received Date:	29 December 2017
Revised Date:	13 March 2018
Accepted Date:	08 May 2018

Please cite this article as: Daniel B. Shank, Alyssa DeSanti, Attributions of Morality and Mind to Artificial Intelligence after Real-World Moral Violations, *Computers in Human Behavior* (2018), doi: 10.1016/j.chb.2018.05.014

Attributions of Morality and Mind to Artificial Intelligence after Real-World Moral Violations

Daniel B. Shank*, Alyssa DeSanti

Department of Psychological Science, Missouri University of Science and Technology, 500 W. 14th Street, Rolla, MO 65409 USA

The media has portrayed certain artificial intelligence (AI) software as committing moral violations such as the AI judge of a human beauty contest being "racist" when it selected predominately light-skinned winners. We examine people's attributions of morality for seven such real-world events that were first publicized in the media, experimentally manipulating the occurrence of a violation and the inclusion of information about the AI's algorithm. Both the presence of the moral violation and the information about the AI's algorithm increase participant's reporting of a moral violation occurring in the event. However, even in the violation outcome conditions only 43.5 percent of the participants reported that they were sure that a moral violation occurred. Addressing whether the AI is blamed for the moral violation we found that people attributed increased wrongness to the AI – but not to the organization, programmer, or users – after a moral violation. In addition to moral wrongness, the AI was attributed moderate levels of awareness, intentionality, justification, and responsibility for the violation outcome. Finally, the inclusion of the algorithm information marginally increased perceptions of the AI having mind, and perceived mind was positively related to attributions of intentionality and wrongness to the AI.

Keywords: artificial intelligence; morality; responsibility; algorithm; attributions; perceived mind

* Corresponding author. *500 W. 14th Street, H-SS Building, Department of Psychological Science, Missouri University of Science and Technology, Rolla, MO 65409, USA. Tel: 573-341-4823. E-mail: shankd@mst.edu.*