May Machines Take Lives to Save Lives?
Human Perceptions of Autonomous Robots (with the Capacity to Kill)

Matthias Scheutz and Bertram Malle
Tufts University and Brown University
matthias.scheutz@tufts.edu and bertram_malle@brown.edu

The possibility of developing and deploying autonomous "killer robots" has occupied news stories for quite some time, and it is also increasingly discussed in academic circles, by roboticists, philosophers, and lawyers alike. However, surprisingly little work has focused on investigating human perceptions of using lethal force in autonomous robots, i.e., whether and when humans would find it acceptable for autonomous robots to use lethal force, in military contexts, and beyond.

In this paper, we report first results from an empirical study designed to investigate ordinary people's moral expectations and judgments about an autonomous robot that must decide whether to kill some human lives to save others. Specifically, we conducted an online experiment using a variant of the well-known Trolley dilemma in which we compared people's evaluations of both human and robot moral decision-making. This design permits us to pinpoint where ordinary moral expectations are the same for humans and robots and where they are different. The results can then inform functional, moral, and legal requirements for autonomous robots that have the capacity to take or save lives – requirements that such robots must meet for their actions to be (maximally) acceptable to humans.