Too familiar to ignore, but too different to tolerate: The role of the law in regulating emotional expression by robots to mitigate the harmful effects of robot deception

N. Gunawardena

Department of Private and Comparative Law, Faculty of Law, University of Colombo, Sri Lanka

A long-standing argument for why robots will never "take over" and pose an existential threat to humans is that we will always have a "kill switch" to deactivate them. But what if a robot manipulates us into not flipping that switch? Recent studies show that robots are capable of deception. Despite ongoing debates regarding the specific nature of this capability, such as whether it is indicative of a theory of mind or consciousness, robot deception (RD) can be "intentional" from at least a behaviorist point of view. The expression/mimicking of emotions is one of the most effective ways in which robots engage in deception. This poses unique ethical and regulatory challenges because it is not ostensibly malicious or threatening. We are already experiencing the harmful effects of RD in our daily lives, particularly the risks associated with over-reliance. However, RD can also be beneficial in certain use cases. This paper addresses the question of how emotional expression by robots should be regulated to mitigate the harmful effects of robot deception. The EU Artificial Intelligence Act (AIA) currently establishes the most comprehensive framework for the regulation of robots and provides a starting point to assess the efficacy of contemporary regulatory standards. Thus, this paper will involve a conceptual/normative analysis of RD combined with a doctrinal analysis of the AIA. Section 2 begins by conceptualizing RD; analyzing its nature, how it manipulates humans, and the harm it can cause. Section 3 evaluates the applicable provisions of the AIA and reveals that the harm caused by expressive robots is being underestimated. Therefore, Section 4 proposes expanding the definition of "harm" to include long-term structural harm and classifying robots based on their different deceptive capabilities. This paper also considers the possibility of imposing design-related restrictions to "trap" expressive robots within the uncanny valley.

Keywords: *Robot deception, Emotional expression by robots, Artificial Intelligence, EU AI Act, Over-reliance on AI*