The Trolley Problem and its Consequences

Nina Papenfuß, Freie Universität Berlin

Abstract: Originally being one of the core fields of Moral Philosophy, moral dilemmas are gaining new attention in the field of artificial intelligence. For instance, when wanting to develop and deploy autonomous cars one is ultimately confronted with the so called trolley problem. This paper introduces the trolley problem and its theoretical and practical consequences. First, the trolley problem will be examined from a philosophical perspective, treating it as a thought experiment and formulating some resulting ethical principles. Second, it will be argued that thinking about how to deal with the trolley dilemma is of great practical importance in order to not only keep traffic as save as it is, but even increase its safety. In spite the fact that the *ideal* trolley problem rarely occurs, some variation of it – the so-called *statistical* trolley problem – is of high relevance for autonomous vehicles, regardless if applied to an implicit or explicit moral agent. Third, recent empirical studies on the "public morality" of the trolley problem – such as the moral machine experiment - and the resulting social dilemmas will be examined. The methodology of the experiment itself bears the risk of social subjectivism as well as cultural relativism. Consequently, it will be argued, empirical studies on public morality cannot substitute a democratic debate and a democratic decision making about what factors to include in the deployment and regulation of ethical autonomous vehicles.

References

Arnold, T, Scheutz, M.: Against the moral Turing test: Accountable design and the moral reasoning of autonomous systems. In: Ethics and Information Technology 18, pp. 103-115 (2016).

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon J.-F., Rahwan, I.: The Moral Machine Experiment. In: Nature 563, pp. 59-64 (2018).

Bonnefon, J.-F., Shariff, A., Rahwan, I.: The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars. In: Proceedings of the IEEE 107, pp. 502-504 (2019).

Bonnefon, J.-F. Shariff, A., Rahwan, I.: The social dilemma of autonomous vehicles. In: Science 352, 1573-1576 (2016).

Crawford, K., Calo, R.: There is a blind spot in AI research. In: Nature 538, pp.311-313 (2016).

Cushman, F., Young, L.: The Psychology of Dilemmas and the Philosophy of Morality. In: Ethical Theory and Moral Practice 12, pp. 9-24 (2009).

Goodall, N. J.: Away from trolley problems and toward risk management. In: Applied Artificial Intelligence 30, pp. 810-821 (2016).

Goodall, N. J.: Ethical decision making during automated vehicle crashes. In: Transportation Research Record 2424, pp. 58-65 (2014).

Greene, D., Hoffmann, A. L., Stark, L.: Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning. In: Proc. 52nd Hawaii International Conference on System Sciences, pp. 2122-2131 (2019).

Grunwald, A. (Ed.): Handbuch Technikethik. J.B. Metzler, Weimar, Stuttgart (2013).

Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. In: Nature Machine Intelligence 1, pp. 389-399 (2019).

Maurer, M., Gerdes, J. C., Lenz, B., Winner, H.: Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte. Springer Vieweg, Wiesbaden (2015).

Moor, J.H.: The nature, importance, and difficulty of machine ethics. In: IEEE Intelligent Systems 21, pp. 18-21 (2006).

Rath, M., Krotz, F., Karmasin, M. (Ed.): Maschinenethik: Normative Grenzen autonomer Systeme. Springer, Berlin (2019).

Shariff, A., Bonnefon, J.-F., Rahwan, I.: Psychological roadblocks to the adoption of self-driving vehicles. In: Nature Human Behavior 1, pp. 694-696 (2017).

Thomson, J. J.: The trolley problem. In: Yale Law Journal 94, pp. 1395-1415 (1985).

Wittpahl, V. (Ed.): Künstliche Intelligenz: Technologie, Anwendung, Gesellschaft. Springer, Berlin (2019).

Weizenbaum, J.: Die Macht der Computer und die Ohnmacht der Vernunft. Suhrkamp, Frankfurt am Main (1978).

Winfield, A.F., Michael, K., Pitt, J., Evers, V.: Machine Ethics: The Design and Gouvernance of Ethical AI and Autonomous Systems. In: Proceedings of the IEEE 107, pp. 509-517 (2019).

Woopen, C., Jannes, M. (Ed.): Roboter in der Gesellschaft: Technische Möglichkeiten und menschliche Verantwortung. Springer, Berlin (2019).