
CHAPTER 8

# Automation and Ethics

**Abstract** Decision-makers in business can expect to face a range of new ethical challenges connected to automation and digitalization. One notable example is that of the programming of self-driving cars. It is likely that these cars can contribute to considerably safer traffic and fewer accidents, as these vehicles will be able to respond much faster and more reliably than fallible human drivers. However, they also raise ethical questions about how to prioritize human lives in situations where either people inside or outside the car will die. Here the reflections are similar to those we have encountered with regard to the trolley problem. Another set of ethical challenges arise in connection with automation and employment. Companies will be in a position to automate processes that have previously been handled by humans, with the aim of cutting costs and enhancing product quality. It will also make current employees redundant. This chapter introduces one conceptual distinction relevant to keep track of automation and ethics between proscriptive and prescriptive ethics, or between avoid-harm ethics and do-good ethics.

**Keywords** Automation • Artificial intelligence • Self-driving cars • Proscriptive ethics • Prescriptive ethics

What should autonomous, self-driving cars be programmed to do in a situation where five children have entered the road in front of the car and



Ø. Kvalnes, *Moral Reasoning at Work*,
https://doi.org/10.1007/978-3-030-15191-1_8

the choice is continuing straight ahead, killing the children, or steering the vehicle out of the road and into a concrete wall, with the result that the car is damaged and the one person inside it dies? These are the kinds of situations that car manufacturers and their programmers are pondering as they are paving the way for a time where humans will not need a driver's license, as the vehicles will drive themselves (Borenstein, Herkert, & Miller, 2017; Gogoll & Müller, 2017). There is a striking similarity to the trolley problems discussed in Chap. 3. Even here it is a question of how to prioritize lives, as well as whether one should follow the utilitarian doctrine of maximizing utility or bring in duty ethics considerations (Nyholm & Smids, 2016).

One study has found that a majority believe the cars should be programmed according to utilitarian ethics (Bonnefon, Shariff, & Rahwan, 2016). In other words, they should take all stakeholders into consideration and aim to minimize human suffering. In the case above, then, the car should sacrifice the one person inside it in order to save the lives of the five children.

A director of Mercedes-Benz made a statement that was interpreted to mean that the self-driving cars his company produces would not follow this pattern. Instead, they would give priority to the lives of those inside the car, every time. This statement had to be retracted, as car companies are not allowed to make these kinds of principled life-and-death decisions (Vijayenthiran, 2016). It is for the authorities to decide whether the cars should be made into utilitarian machines. However, internationally the situation is that authorities so far are silent on this issue. It remains to be seen whether there will be global agreement on how the cars should be programmed or whether there will be local differences.

The study by Bonnefon et al. (2016) points in the direction of a utilitarian solution to the ethical challenge of programming driverless cars but also identifies a paradox for this normative theory. Participants in the study were also asked what kind of car they themselves would buy, and here the majority answered that they would avoid getting a car programmed in the utilitarian way but would rather have one that promised security to those inside it, even when they are in the minority compared to those in danger outside it. So even though the majority thought that it made good ethical sense to minimize human suffering by saving the five children by sacrificing the person inside the vehicle, that was not the kind of car they would purchase. The authors point to a paradox for the utilitarian theory: From the perspective of maximizing utility, the introduction of self-driving cars

is a phenomenally good thing, as it is likely to lead to a drastic reduction in traffic accidents. The sooner such a shift happens, the better, as while we are waiting, more people will die and get injured in traffic. However, if the cars are programmed according to utilitarian ethics, people are much less likely to shift to that kind of vehicle, so the transition will happen much more slowly, and utility will be lost. Paradoxically, then, the utility of safer traffic appears to depend on programming that goes against utilitarian thinking.

The tempo of the transition to automated solutions is also likely to be affected by the ways in which liability issues are treated. Who do we hold accountable if things go wrong in automated processes? When decision-makers in companies consider artificial intelligence solutions, all six questions in the Navigation Wheel are relevant, and the question about legality may be particularly difficult to answer, as there are few previous practices to compare the current situation with. There is limited legal tradition or precedent to appeal to. Researchers have started to address alternative models for distributing responsibility after accidents involving bots (Abbott, 2017; Headrick, 2014; Kessel & von Bodungen, 2018). One suggestion has been that as machines have the potential to significantly reduce the risk of accidents, the legal framework should encourage automation and protect the manufacturers against strict liability charges (Abbott, 2017). If the companies developing automated solutions face the risk of being held fully responsible for any bad outcome in the bots' behaviors, it may make them hesitant and slow in introducing those solutions. From a utilitarian point of view, this would be unfortunate, as the introduction of automation can improve the quality of services and make traffic and other potentially harmful activities safer. One way forward can be to introduce an alternative way of thinking about negligence and responsibility for bad outcomes, where the standard shifts from being based on what a hypothetical reasonable person would have done to what a hypothetical reasonable computer would have done (Abbott, 2017).

Automation raises ethical questions in a range of business areas. In finance, the use of autonomous trading agents is already prevalent, and with that activity come ethical questions that have still not been adequately addressed (Wellman & Rajan, 2017). High-frequency trading in stock markets occurs between bots, with hardly any human intervention. Davis, Kumiega, and Van Vliet (2013) argue that current disciplinary standards do not adequately deal with the ethical problems generated by these procedures and claim that the financial industry needs a cross-cultural ethical

framework to address them. The current system is vulnerable, and both the regulators and the industry itself need to identify principles for reasonable distribution of risk and responsibilities (Davis et al., 2013).

Automation also raises ethical questions in the realm of employment. Mechanical minds are already outperforming humans in a range of activities, and this tendency is on the increase. Companies can cut costs and improve the quality of their products by introducing artificial intelligence solutions. Researchers disagree on the severity of the threat to human employment and the likely speed of the development. One pessimistic view is that as much as 47% of the current jobs in the United States are under high risk of being replaced by bots (Frey & Osborne, 2017), whereas more optimistic scenarios assume that automation will generate new jobs for humans to become involved in (Autor, 2015; Nokelainen, Nevalainen, & Niemi, 2018). The future for established professionals such as lawyers, auditors, doctors, and others with specialized knowledge is also open, as research points to the likelihood that the need for their traditional services will decrease due to rapid advances in automation (Laster, 2016; Susskind & Susskind, 2015, 2016).

One overarching ethical challenge for developers and users of automated systems is how to implement ethically sound decision-making procedures (Wallach & Allen, 2008). Artificial intelligence can absorb and use vastly more information than human beings are capable of at a dramatically higher speed. Earlier in this book we assumed that ethical decision-making is an example of Kahneman's System 2, where we slow down the tempo, in order to take the relevant factors into careful consideration. This is not so with the ethical decision-making of mechanical minds, with their vastly superior ability to handle information quickly. The difference in tempo aside, the automated decisions must be based on reasonable ethical principles and norms. How can we incorporate ethics into the complex algorithms and procedures that mechanical minds or computers perform?

Ethical principles can be integrated into the artificial intelligence through a bottom-up procedure, where the bot is designed to register and act in accordance with the aggregate moral convictions and beliefs it somehow encounters and registers in the society in which it operates. Alternatively, ethical principles can be programmed into the bot in a top-down process, where programmers and engineers dictate the content based on specific legal and regulatory boundaries (Allen, Smit, & Wallach, 2005; Allen, Varner, & Zinser, 2000; Baum, 2017; Etzioni & Etzioni, 2017; Wallach & Allen, 2008; Wallach, Allen, & Smit, 2008).

The bottom-up approach assumes that a bot can gradually learn ethics and integrate moral standards through interactions in a social environment (Allen et al., 2000, 2005; Wallach & Allen, 2008; Wallach et al., 2008). It can register information about what counts as good or bad and right or wrong behavior from observations of how people behave, as well as how they respond favorably or unfavorably to each other's actions. The ethical principles and moral standards or convictions of the bot can be updated and revised regularly after it has interacted with and learned from others. One challenge for this approach became evident with the launch of Microsoft's chat bot Tay, which was supposed to learn ethical principles and standards for morally acceptable behavior through communication with humans. Tay quickly started to speak vulgarities, even though this was not intended or wished for most of the people who interacted with it. Tay's training was dominated by a vocal minority who used vulgar language in its repetitive interactions with it (Baum, 2017). The bot needs to interact with the right people and in the right manner in order to integrate the right set of moral standards and behaviors, and it remains a challenge to establish the proper quality controls.

Etzioni and Etzioni (2017) argue that it is both impossible and unnecessary to implant ethics into bots. They reject both the bottom-up and the top-down approach. Instead they call for societies and authorities to set legal limits on what the machines are allowed to do. The scope of action for bots should in this view be regulated by the collective, democratic processes of lawmaking and regulations and would not leave room for ethical considerations to be taken by the machines themselves. There might be technological challenges in making the machines comply with the regulations, but that should not be confused with the task of making them into autonomous decision-makers operating from their own moral standards or ethical principles. A similar view has been expressed by Yampolskiy (2013) who rejects the idea that machines can be programmed to make ethical decisions. The primary decision-makers are the lawmakers who should decide the scope of action for engineers and programmers, whose role it is to develop safe and reliable engineering solutions. In this view, automated operations should be dictated so as to be in compliance with laws and regulations, and those should in turn be in harmony with the society's moral standards.

Reflections on the ethics of automation tend to focus on the dangers and threats of technological advances that create intelligences capable of

outperforming human beings. It is worth noting that artificial intelligence can also make positive contributions, even when studied from an ethical perspective. The distinction between proscriptive and prescriptive ethics (Janoff-Bulman, Sheikh, & Hepp, 2009) is useful to bring out the full scope of the ethical dimension of automation.

Proscriptive ethics can also be called avoid-harm ethics, which brings attention to the possible pitfalls of behaviors and decisions. In the context of automation, it is an ethics that warns us against mass unemployment, lack of control over decision-making procedures, and the scary scenario where the bots are smarter than humans and begin to communicate with each other in ways incomprehensible to human beings.

Prescriptive ethics can also fall under the name of do-good ethics and concerns itself with how behaviors and decisions can improve and advance human conditions. It is important to keep in mind that there is a prescriptive dimension to the ethics of automation, in that bots can improve the services available to human beings through safer traffic; higher quality and precision in medicine; improved control over health, security, and environment issues in workplaces; and so on. It is not that research on automation has neglected the positive aspects, but it has chosen to place it outside the scope of ethics. My suggestion here is that the ways in which automation can potentially promote well-being for humans warrant an inclusion under the ethics heading, or more precisely as material for prescriptive ethics.

In an ongoing research project on automation and ethics, Miha Škerlavaj, Ketill Berg Magnússon, and I have asked EMBA students in Norway and Iceland about their perceptions and expectations in this area. These are students who already have extensive business experience. Most of them have already encountered automated solutions and mechanical minds in their jobs and point to how they enhance quality and efficiency, potentially disrupt employment structures, and potentially increase the gap between rich and poor populations, thus creating social tensions.

One surprising finding in our material is that several students point to how mechanical minds can reduce bribery, corruption, and other morally questionable behaviors in business:

> I consider that one of the most important advantages of electronic purchasing platforms is that they eliminate the risk of bribes or other forms of corruption to influence a decision after a bidding process.
>
> Student A, Oslo

> All this effort has the aim of increasing automation with new IT systems. The purpose is to meet growth and limit hiring as much as possible, increase service, and minimize fraud.
>
> Student B, Reykjavik

> The plus for the banking business will be the benefits from lower costs with the underwriting department and less fraud losses due to no judgement or human intervention in the process.
>
> Student C, Oslo

The common assumption in these claims is that automation can reduce the dependence on human interactions and social arenas where corruption currently takes place. It is not that the bots come along armed with a superior morality but rather that they can be programmed into sticking to the facts and figures and not be influenced by ingratiating behaviors or attempts to gain improper advantages through the use of improper business methods.

In this chapter, we have seen that when we study automation through the lens of ethics, it has a prescriptive, do-good dimension and a proscriptive, avoid-harm dimension. The emergence of artificial intelligence and bots in organizational settings introduces possibilities that transcend our current capacities for understanding. With this development come ethical challenges for decision-makers. The programming of autonomous vehicles has already received plenty of attention, and other issues will follow. To some extent, traditional ethical theories such as utilitarianism and duty ethics offer guidelines on how we can reason and reflect about those choices, but a richer set of concepts may be called for in order to keep track of developments in this area.

## References

Abbott, R. (2017). The reasonable computer: Disrupting the paradigm of tort liability. *George Washington Law Review, 86*.

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology, 7*(3), 149–155.

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence, 12*(3), 251–261.

Autor, D. H. (2015). Why are there still so many jobs? The history and future of workplace automation. *The Journal of Economic Perspectives, 29*(3), 3–30.

Baum, S. D. (2017). Social choice ethics in artificial intelligence. *AI & SOCIETY*. https://doi.org/10.1007/s00146-017-0760-1

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science, 352*(6293), 1573–1576.

Borenstein, J., Herkert, J., & Miller, K. (2017). Self-driving cars: Ethical responsibilities of design engineers. *IEEE Technology and Society Magazine, 36*(2), 67–75.

Davis, M., Kumiega, A., & Van Vliet, B. (2013). Ethics, finance, and automation: A preliminary survey of problems in high frequency trading. *Science and engineering ethics, 19*(3), 851–874.

Etzioni, A., & Etzioni, O. (2017). Incorporating ethics into artificial intelligence. *The Journal of Ethics, 21*(4), 403–418.

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change, 114*, 254–280.

Gogoll, J., & Müller, J. F. (2017). Autonomous cars: In favor of a mandatory ethics setting. *Science and Engineering Ethics, 23*(3), 681–700.

Headrick, D. (2014). The Ethics and Law of Robots. *Research Technology Management, 57*(3), 6–7.

Janoff-Bulman, R., Sheikh, S., & Hepp, S. (2009). Proscriptive versus prescriptive morality: Two faces of moral regulation. *Journal of Personality and Social Psychology, 96*(3), 521.

Kessel, C., & von Bodungen, B. (2018). Germany's new road traffic law—Legal risks and ramifications for the design of human-machine interaction in automated vehicles. In *Advanced Microsystems for Automotive Applications 2017* (pp. 227–236): Springer.

Laster, K. (2016). Future of the law: Doomsday prophet or optimist?: Susskind predicts the end of the professions. *Bulletin (Law Society of South Australia), 38*(8), 28.

Nokelainen, P., Nevalainen, T., & Niemi, K. (2018). Mind or machine? Opportunities and limits of automation. In C. Harteis (Ed.), *The impact of digitalization in the workplace: An educational view* (pp. 13–24). Cham: Springer International Publishing.

Nyholm, S., & Smids, J. (2016). The ethics of accident-algorithms for self-driving cars: An applied trolley problem? *Ethical Theory and Moral Practice, 19*(5), 1275–1289.

Susskind, R., & Susskind, D. (2015). *The future of the professions: How technology will transform the work of human experts.* USA: Oxford University Press.

Susskind, R., & Susskind, D. (2016). Technology will replace many doctors, lawyers, and other professionals. *Harvard Business Review, 11*.

Vijayenthiran, V. (2016, 18.10). Mercedes is backtracking on claims its self-driving cars will kill pedestrians over passengers in close calls. *Business Insider*.

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. New York: Oxford University Press.

Wallach, W., Allen, C., & Smit, I. (2008). Machine morality: Bottom-up and top-down approaches for modelling human moral faculties. *AI & SOCIETY, 22*(4), 565–582.

Wellman, M. P., & Rajan, U. (2017). Ethical issues for autonomous trading agents. *Minds and Machines, 27*(4), 609–624.

Yampolskiy, R. V. (2013). Artificial intelligence safety engineering: Why machine ethics is a wrong approach. In *Philosophy and Theory of Artificial Intelligence* (pp. 389–396). Springer.