5.5 Implementing Ethical Deliberation

Most current work in AI ethics is about developing the algorithms that enable an AI system to consider the ethical aspects of its decisions. In reality the spectrum of possibilities to implement decision-making is much wider. In many cases, an AI system that is fully and solely in charge of determining the best ethical behaviour is not only too complex to implement, but also not necessary. Humans have long learnt to ask for help from others and to build their environments to facilitate a certain type of behaviour. This is, for instance, the role of laws and social norms. It is thus only logical to expect the same possibilities to be available to AI systems.

In the following, we identify four possible approaches to design decision-making mechanisms for autonomous systems and indicate how these can be used for moral reasoning by AI systems.

- **Algorithmic:** this has the aim to incorporate moral reasoning fully in the system's deliberation mechanisms. According to [132], an AI system can autonomously evaluate the moral and societal consequences of its decisions and use this evaluation in its decision-making process. Here 'moral' refers to principles regarding *right* and *wrong*, and 'explanation' refers to algorithmic mechanisms to provide a qualitative understanding of the relationship between the system's beliefs and its decisions. This approach requires complex decision-making algorithms, as we have discussed in this chapter, but also requires the system to be able to do this reasoning in real time.
- **Human in command:** in this case, a person or group of persons take part in the decision process. Different collaboration types can be identified, ranging from that of an autopilot, where the system is in control and the human supervises, to that of a 'guardian angel', where the system supervises human action. From a design perspective, this approach requires the inclusion of means to ensure shared awareness of the situation, such that the person making decisions has enough information at the time she must intervene. Such interactive control systems are also known as human-in-the-loop control systems [84].
- **Regulation:** these are approaches where an ethical decision is incorporated, or constrained, in the systemic infrastructure of the environment. In this case, the environment ensures that the system will never get into moral dilemma situations. That is, the environment is regulated in such a way that deviation is made impossible, and therefore moral decisions by the autonomous system are not needed. This is the mechanism used e.g. in smart highways, linking road vehicles to their physical surroundings, where the road infrastructure controls the vehicles [93]. In this case, ethics are modelled as regulations and constraints in the infrastructure, where AI systems can get by with limited moral reasoning.

- **Random:** finally, we should also consider the situation in which the AI system randomly chooses its course of action when faced with a (moral) decision. The claim here is that if it is ethically problematic to choose between two wrongs, a possible solution is to simply not make a deliberate choice.[3] The Random mechanism can be seen as an approximation of human behaviour, and can be applied to any type of system. Interestingly, there is some empirical evidence that, under time pressure, people tend to choose justice and fairness over careful reasoning [14]. This behaviour could be implemented as a weak form of randomness. Research is need to understand the acceptability of random approaches.

As we saw in Section 5.4.1, who is consulted and how individual values are aggregated also influence which implementation approach is most suitable. Moreover, different societies interpret values differently. Using the value system introduced in Section 3.3, it can be expected that in societies that prioritise Conformity, people will be more likely to choose a regulatory approach as implementation mechanism, where legal norms and institutions take responsibility for the decisions, whereas Egalitarian societies might accept a random mechanism for decision-making, which would make no judgement and express no preference between passengers and pedestrians.

In order to make explicit the values and value priorities of designers and stakeholders, methodologies for Design for Values are needed, as described in Chapter 3.

5.6 Levels of Ethical Behaviour

As AI systems are increasingly able to interact autonomously and to have some awareness of their (social) environment, people are changing their ideas about them. Automated assistance of whatever kind doesn't simply enhance our ability to perform the task; it changes the nature of the task itself as well as how people engage with machines. Even though AI systems are artefacts, people are increasingly starting to see machines not as simple tools but as team members or companions.

According to [132], different levels of ethical behaviour should be expected of each of these categories (Figure 5.3). The simplest are *tools*, such as a hammer or a search engine, which have no or very limited autonomy and social awareness, and are therefore not considered to be ethical systems. Nevertheless, values are incorporated into their design, often implicitly, which leads to different behaviours. The next type of systems, *assistants*, have limited autonomy but are aware of the social environment in which they interact. These systems are expected to have functional morality, meaning that re-

[3] cf. Wired: https://www.wired.com/2014/05/the-robot-car-of-tomorrow-might-just-be-programmed-to-hit-you/