Chapter 4
Trust and Fairness in AI Systems

This chapter discusses the role that trust and fairness play in the acceptance of AI systems. We then relate trust and fairness in AI systems to five principles: Non-maleficence, Beneficence, Autonomy, Justice and Explicability.

There are many contexts in which one can use the word trust. For example, one might ask a person if they trust their spouse, if they trust a company such as Facebook or Google with their data, if they trust the government with their health data, or if they trust their housekeeper or their babysitter.

In human contexts, these questions all involve some kind of vulnerability to harm. Facebook and Google might sell data to some disreputable firm seeking to use it for manipulative purposes. People might believe that the government will hand over their health data to insurers which might raise their premiums. A housekeeper might steal. A babysitter might neglect the children in his or her care. In all of these cases, the person trusting has a vulnerability to some kind of harm or damage. If we do not trust people, we will be on our guard. We will not entrust them with anything important. We might even seek to avoid them completely if we can. Furthermore, uncertainty is a prerequisite for trust because vulnerability is diminished if one knows the outcome upfront.

Lee and See (2004) defined trust as "the attitude that an agent will help achieve an individual's goals in a situation by uncertainty and vulnerability." In the context of robots and AI systems, the question of trust is similar but differs in some respects. When a human trusts another human, the trustor knows that they are making themselves vulnerable to the actions of the other person. When a person trusts an AI agent, it is unclear if the machine is making it's own decision or following some predefined scripted behaviour. Furthermore, people have experiences with other people and have certain expectations about the behaviours, norms and values. We may trust a pilot to safely land an aircraft. When a user is interacting with a robot or an AI system, these experiences may become useless or even misleading. We cannot be

C. Bartneck et al., *An Introduction to Ethics in Robotics and AI*,
SpringerBriefs in Ethics, https://doi.org/10.1007/978-3-030-51110-4_4

certain that a system will act in our best interest. The rest of this chapter discusses trust and fairness in the context of humans trusting machines such as AIs and robots.

4.1 User Acceptance and Trust

Trust is critical to user acceptance. People will avoid using systems they do not trust. Obviously, businesses that make products that are not trusted will not see their products being used—and will thus struggle in the marketplace. For example, companies in the U.S. are obliged to comply with the Patriot Act that empowers the government to access data stored on cloud computers. Customers from Europe might feel uncomfortable about granting the U.S. government such a privilege. After the European Court of Justice overturned the long-standing US-EU Safe Harbor agreements in 2015, several cloud storage companies opened data centers in Europe to regain the trust of their European customers.

Culture influences the trust that people place in AI systems and robots. Haring et al. (2014a) shows that these cultural differences impact how people view robots. Cross-cultural studies also demonstrate differences in positivity (Haring et al. 2014b) which in turn impacts trust. Cultural factors may also influence the extent to which people follow a robot's recommendations (Wang et al. 2010).

In the case of machines, the question of trust can be broken up into "functional" and "ethical" elements.

4.2 Functional Elements of Trust

According to Lee and See (2004), users calibrate their trust in a machine based on a variety of factors including the system's reliability. Hancock et al. (2011) looked at a variety of factors influencing trust in automation. These included performance-based factors such as a system's dependability, false alarm rate, transparency, and task complexity. This work showed that performance and reliability are the dominant factors determining if a person trusts a machine.

People will be wary and suspicious of machines that are potentially harmful or dangerous. The possibility of loss of life or injury tends to cause people to reconsider whether or not to trust a machine.

4.3 Ethical Principles for Trustworthy and Fair AI

The European ethical principles for AI, presented by the AI4People group in 2018 (Floridi et al. 2018), suggest five principles for ethics of AI, which can be tied to trust and fairness in the following way.

4.3.1 Non-maleficence

The principle of Non-maleficence states that AI shall not harm people. AI systems that harm people (and animals or property) are a risk, both for individuals as well as for companies (cf. Chap. 6). In terms of trust and fairness, this extends to issues such as bullying and hate speech which are salient examples of the principle of non-maleficence being violated online.

4.3.1.1 Social Media Bullying and Harassment

The cases of school children being bullied by classmates or employees bullied at workplace are numerous. There are cases where this has even lead to suicide. While such bullying certainly existed prior to the advent of digital technologies, it has, through social networks, acquired a new level of concern, as indicated by surveys.[1] The risk of being exposed to large groups or to the public in general has risen dramatically. Consequently, in a number of countries, laws have been passed against bullying in digital media (Sweden was the first country to do so in 1993).

4.3.1.2 Hate Speech

Hate speech has in recent years gained considerable media coverage. Hate speech is a type of speech that is not just specifically directed against a particular person, but can also be directed against groups or entire parts of a population. It might, for example, attack groups on the basis of their ethnic origin, sexual orientation, religion, gender, identity, disability and others. Already the International Covenant on Civil and Political Rights", in force since 1976, includes a statement according to which "any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law.

Laws against hate speech have since then been passed in a number of countries. For example, Belgium passed a law in 1981 which states that incitements to discrimination, hatred of violence against persons or groups of race, colour, origin or national or ethnic dissent are illegal and subject to punishment according to the Belgian penal code. Other countries like Canada, New Zealand or the United Kingdom have similar laws (which might however vary in detail and punishment).

A new level of hate speech laws was however reached in 2017 in Germany, when a bill was passed by the German Bundestag which specifically criminalises hate speech on social media. The law also insists that social networks (like Facebook or others) may be fined very large sums of up to 50 million EUR—in case they do not actively seek and successfully remove certain content within a week.

The passing of this law was controversial, with a number of German, but also many international commentators, stating that such a law is very far-reaching and

[1] https://www.theguardian.com/uk-news/2017/aug/14/half-uk-girls-bullied-social-media-survey.

will have a number of non-intended and counterproductive consequences. Since then, social networks like Twitter or Facebook have taken many efforts to comply with this new law. And while they have certainly succeeded in removing quite a lot of illegal content (as the European Commission acknowledged in 2019,[2] there have also been many issues with content being removed either by accident or due to over-interpretation of certain statements. Appeal of removal decisions is also important and social networks are starting to implement this.

The balance between hate speech and freedom of speech is being decided within and across nations. Social media has infused these debates with elements of cultural and national perspectives. Large companies such as Facebook and Google are being forced to edit their content. Some may view this as a necessary means for combating racism others may view this as in attack on freedom of speech.

4.3.2 Beneficence

The principle of Beneficence states that AI shall do people good. This is a general principle from bioethics according to which the benefits from a treatment must outweigh the potential harms (Zalta 2003). An AI system needs to consider ethical rules to become trustworthy and fair which will enable it to make life better. Examples of ways in which an AI system may demonstrate beneficence with respect to societal problems are:

- the reduction of fatalities and accidents by autonomous vehicles;
- providing robotic support and care for the ageing society (see Sect. 9.2)
- the use of telemedicine in remote areas;
- smart grid based sustainability improvements;
- improvements in terms of biodiversity, e.g., by AI applications to preserve endangered species
- the use of robots and AI in education, for example by using AI for individualised learning or supporting students outside the classroom.

4.3.3 Autonomy

The principle of Autonomy states that AI shall respect people's goals and wishes. To be sure, autonomy has several meanings, and we refer to some of them in Chaps. 10 and 11. Traditionally in AI and robotics, the term autonomy refers to an AI system's or robot's ability to operate without human intervention. In this section, however, we focus on the ethical principle of autonomy. In the context of bioethics, it usually refers to patients having the right to decide for themselves whether or not to undergo a

[2] http://fortune.com/2019/02/04/facebook-twitter-google-eu-hate-speech/.

treatment (Zalta 2003). This entails giving patients the right to refuse life-saving procedures and to decide not to take risk-reducing medications. For example, instances are known in which people died as a result of refusing blood transfusion for religious reasons. Generally, the courts have found that parents do not have a right to impose their views, such as refusing a blood transfusion, on their children (Woolley 2005). However, once a child becomes an adult, they can refuse treatment.

More generally, autonomy refers to the ability of a person to make decisions. People can decide whether or not they want to take risks to earn more money or have more fun. Sherpas carrying packs for mountaineers climbing Everest can earn five times more money than they can working in the fields of Nepal. However, they run the risk of being caught in an avalanche that can injure or kill them. People should be permitted to assume risks but they should know what risks they are assuming. An AI's respect of human autonomy would support permitting a degree of human self-harm and permitting humans to assume risks that might lead to harm. After all, humans might like to go rock-climbing or ride motorcycles.

There are ethical limits to autonomy. Suppose a child says to a robot, "pull my sister's hair, she's been mean to me and I hate her!" In this case, should the robot obey the child? If an adult says, "smack my daughter, she's been naughty" should the robot obey? Generally speaking, systems should not help people pursue illegal or immoral goals. Moreover, systems should not be used to perform illegal or immoral acts. There is no strong case for systems allowing users to employ the systems to harm others unless there is good reason. One should also be very careful when considering to design systems that permit machines to harm humans that ask to be harmed. For example, one might decline to develop an euthanasia robot. There is the risk such a robot might not sense mental illness. Also, one might hesitate to delegate such grave decisions to machines at all.

However, there are some cases in which robots are designed to use force against humans. The police may need to harm people in some cases such as apprehending violent offenders. For example, in 2016, the Dallas Police Chief ordered the use of a tele-operated robot to carry an explosive to a perpetrator who had shot 10 officers, killing 5 and wounding 5 (Thielmann 2016). The explosion killed the offender. Humans were in control of this particular robot. In the military domain, robots capable of violent action against humans have existed for some time. However, apart from lawful violence carried out by the state, there are relatively few cases where people are comfortable with robots and AIs being designed to harm humans.

Moral philosophy is directly linked to autonomy. If a person does not have autonomy or free will, then it can be argued that this person does not have moral responsibility either. From the deontological point of view the connection between autonomy and ethics can best be demonstrated by Immanuel Kant's moral theory (Kant 1785) which consists of three main formulations:

1. Act only according to that maxim you can at the same time will as a universal law without contradiction.

2. Act in such a way that you treat humanity, whether in your own person or in the person of any other, never merely as a means to an end, but always at the same time as an end.
3. The third practical principle follows from the first two as the ultimate condition of their harmony with practical reason: the idea of the will of every rational being as a universally legislating will.

The formula of universal law tells us that to morally justify an action, one must be able to universalise the moral rule or "maxim." One way to think about this is to pose the question: what if everybody followed this rule? What would the world look like then?

The principle of humanity tells us we should respect the ends (goals, wishes) of other persons. We should not treat other human beings as "mere means" alone. Rather we must consider the ends they have as well as our own.

Kant also formulated his moral philosophy in the so-called "Kingdom of Ends" version of the Imperative which states that: "Act according to maxims of a universally legislating member of a merely possible kingdom of ends." The principle of autonomy and the Kingdom of Ends formulation tell us that we must walk our own moral talk. That is, we are bound to obey the moral rules we expect others to follow. As autonomous beings, Kant holds, we are obliged to rationally consider our moral standards. Simply following the law is not good enough. It follows that an artificial intelligent system must have autonomy in the Kantian sense to be able to act in an ethical way.

Many writers bundle moral responsibility and moral decision making together in their definitions of what an "ethical agent" is. Some separate these two notions holding that an AI can make moral decisions without being responsible for such decisions. (Welsh 2018). On Kant's conception, a system that is programmed to simply follow rules such as Asimov's Three Laws of Robotics would not be considered an ethical agent.

Isaac Asimov (2 January 1920–6 April 1992) proposed three rules of robotics that would safeguard humanity from malevolent robots.

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

While his work is highly visible in the public media, it has been criticised by philosophers. Asimov eventually added a zeroth law:

0. A robot may not harm humanity, or, by inaction, allow humanity to come to harm.

4.3.4 Justice

The principle of justice states that AI shall act in a just and unbiased way. Justice is often illustrated by a statue of the Roman Goddess Justitia. Frequently, she is depicted with a sword, scales and a blindfold (see Fig. 4.1). The blindfold represents impartiality. The scales represent the weighing of evidence. The sword represents punishment.

Defining "justice" at the human level is a substantial challenge for AI. The main problem for AI is that moral theory is vigorously contested. There are a variety of "sects and schools" that have been engaged in "vigorous warfare" with each other since the "young Socrates debated the old Protagoras" as Mill (1863) puts it. Polling done by Bourget and Chalmers (2014) shows that none of the major schools of ethical theory enjoy firm majority support. Around a quarter of philosophers "accept" or "lean towards" either deontology or consequentialism. Around a third accept or lean towards virtue ethics. Thus generally defining "justice" or "ethics" in terms of what machines can process is hard. There is no agreement on moral theory.

Humans, by contrast, come with "moral intuition" and have learned some things about right and wrong over a number of years. It is often said by some that human

Fig. 4.1 Justitia (*Source* Waugsberg)

moral intuition is a "black box" of "inscrutable" biological code. We do not fully understand how humans make moral decisions. We do not even understand how human brains store information. However, while there is little agreement on moral theory, when it comes to moral practice, there are many actions that are generally considered to be "right" and many that are considered to be "wrong." While ethical controversies rage on topics such as abortion, euthanasia, civil disobedience and capital punishment, there are many moral problems that are far less difficult.

If the scope of application is reduced and the information on which the moral decisions are made can be obtained, then it is possible for AI to make some very specific moral decisions. Often, there are clear regulations and laws that in very specific applications can be taken as normatively definitive. In practice, AI has already led to a number of implemented applications that may have moral and ethical implications.

4.3.4.1 Determining Creditworthiness

Banks and other credit institutions are already, in a number of countries, using AI systems to pre-sort credit applications on the basis of the data available about the applicant. This certainly has a number of advantages, one of which is to be able to come to a decision more quickly and on the basis of more information, making it more rational in theory. However, this may also entail disadvantages, in particular leading to certain biases. For example, the personal information of a credit applicant will in most cases contain information about their neighbourhood. On that basis, and making use of publicly available (or privately collected) further data, a systematic bias against persons from certain residential areas may occur.

An in-depth examination of racial bias related to the use of algorithms for high-risk care management demonstrates many of the challenges and issues associated with the merged concepts of justice, algorithmic fairness, and accuracy (Obermeyer et al. 2019). Researchers examined the inputs, outputs, and objective function to algorithm used to identify patients as high risk of needing acute care and thereby influenced the treatment of millions of Americans. Even though the algorithm specifically excludes race from consideration, the system reasonably uses information about healthcare costs to predict healthcare need. The system uses machine learning to create a model to predict future healthcare costs. It assumes that those individuals that will have the highest healthcare costs will be the same individuals that need the most healthcare, a very reasonable assumption. Yet, this assumption introduces disparities that end up correlating to race. For example, poor patients face greater challenge accessing healthcare because they may lack access to transportation, childcare, or have competing work related demands. They conclude that the central issue is problem formulation: the challenge of developing precise computational algorithms that operate on amorphous concepts. Inevitably the types of precise measures needed for such algorithms include distortions that often reflect structural inequalities and related factors. These issues may be endemic among many industry algorithms across many industries.

4.3.4.2 Use in Court

Depending on the country, AI software is being used in courts. One relatively simple example is the use to determine the order in which cases are brought up to a judge, making use of information on the severity of cases, prior convictions, and more, in order to make a court's work more efficient (Lin et al. 2019). A more impactful use is in supporting judges to determine whether an inmate gets released on probation or not. A 2016 study by ProPublica found that the COMPAS system exhibited a systematic bias against African-American defendants in Broward County, Florida in terms of assess recidivism risk (Angwin et al. 2016).

This case generated considerable controversy. The developers of COMPAS (Northpointe, now Equivant) in their response to ProPublica's statistical analysis argued that ProPublica had "focused on classification statistics that did not take into account the different base rates of recidivism for blacks and whites" (Dieterich et al. 2016). This is a highly technical argument. Several commentators have observed there are multiple conceptions of "fairness" in the machine learning literature. With particular reference to the Broward County recidivism rates it has been demonstrated mathematically that "an instrument that satisfies predictive parity cannot have equal false positive and negative rates across groups when the recidivism prevalence differs across those groups" (Chouldechova 2017).

In many of these cases, countermeasures are not easily devised. There are multiple conceptions of fairness which might conflict. One might be faced with a possible trade-off between fairness, on the one hand, and accuracy on the other. In fact, some argue that efforts to 'blind' algorithms to objectionable categorical information, such as race, may be harmful and a better approach would be to alter how we use machine learning and AI (Kleinberg et al. 2018). While one certainly wants to avoid systematic biases in their AI programming, data should also reflect the actual situation, i.e. "what is the case", otherwise these data will not be very useful for drawing further conclusions and acting on their basis.

It is critical to appreciate the limits of classifications based on statistical models. While much of the Broward County recidivism controversy centred on the difference between false positives between blacks and whites, one study found the underlying predictive accuracy of COMPAS was only around 65%. The authors observed:"these predictions are not as accurate as we might want, particularly from the point of view of a defendant whose future lies in the balance" (Dressel and Farid 2018).

4.3.5 Explicability

The principle of Explicability states that it shall be possible to explain why an AI system arrived at a certain conclusion or result.

Explicability is not to be equated with transparency. While some call for maximal transparency of programs and codes, when it comes to AI systems this might not solve a problem and might even create new problems. Suppose software containing millions of lines of code is made transparent, what would be the benefit of this? First, the software would probably not be intelligible to non-experts, and even experts would struggle with what it means. Second, maximal transparency of software might create a risk vis-a-vis competitors, and hinder further investment in this sector. Due to considerations like these, some parts of the discussion have switched to the concept of "explicability".

Explicability, as Floridi et al. (2018) define it, means both intelligibility and accountability. It is desirable in moral applications that those using AI systems or whose interests are affected by AI systems can "make sense" of the precise way in which an AI made a particular decision. Intelligibility means that the workings of the AI can be understood by a human. The system is not a mysterious "black box" whose internals are unintelligible. Someone, even if only an experienced programmer, can understand the system's workings and explain it to judges, juries and users.

The European Union implemented a legal requirement for the "right to information" (originally called the "right to explanation") within the framework of the General Data Protection Regulation. This holds that people whose interests have been affected by an algorithmic decision have to right to have the algorithm explained the decision to them. Obviously, this poses a challenge for some "inscrutable" machine learning methods such as neural networks. There are some who are untroubled by "inscrutability" in machine learning. They take the view that they can test the system empirically. So long as it works in practice, they do not care if they cannot understand or explain how it works in practice (Weinberger 2018).

For many machine learning applications this may be fine. Yet, in morally charged situations that might be subject to judicial review, there will be a requirement to explain the decision. There may also be a requirement to justify the decision. A key element of moral functioning is not just doing the right thing but justifying what the agent did is right. Moral justification cannot be based on an "inscrutable" black box. However, there is ongoing research into "explainable AI" which seeks to generate explanations for why neural networks make the decisions they make (Wachter et al. 2017). It may be that such research eventually enables machine learning to generate adequate explanations for decisions it makes.

Even so, in practical terms, the use of the COMPAS system to assess recidivism risk and thus affect prospects for probation has been tested in court. In the case of Loomis versus Wisconsin, the plaintiff argued that he was denied "due process" because of the proprietary nature of the COMPAS algorithm meant that his defence could not challenge the scientific basis on which his score was calculated. However, his appeals failed. The judges held that sentencing decisions were not entirely based on COMPAS risk scores. Judges could not base sentences on risk scores alone but could consider such scores along with other factors in making their assessment of recidivism risk.

Accountability at its most basic can be provided in the form of log files. For example, commercial airlines are required to have a flight recorder that can survive a crash.

In the event of a plane crash, the flight recorders can be recovered. They facilitate the investigation of the crash by the authorities. Flight recorders are sometimes called "black boxes". The log files generated by the system that can explain why it did what it did. This data is often used to retrace the steps the system took to a root cause and, once this root cause if found, assign blame.

4.4 Conclusion

To sum up, given a free choice, users will accept systems if they trust them, find them useful and can afford them. Businesses therefore have an incentive to make systems people trust. Trust involves a very complex cluster of concepts. To trust a system users will need to be confident that the system will do good to them or for them. That is it will display beneficence towards them. They will need to be confident the system will not do bad things to them such as harm them or steal from them or damage their interests in some other way (e.g. breach their privacy or cause them embarrassment).

They need to be confident the AI will not compromise their autonomy. This is not to say robots and AIs will never say no to humans. It is merely to say they will only say "no" if there is a good reason. For example, if the human wants to do something wrong with the robot, a morally intelligent robot could (in theory) refuse to obey the human.

People will need to be confident that the system can act justly within its functional scope. If there are clear laws and regulations and no moral controversies this is much easier to implement. Currently, due to the lack of agreement on moral theory, such functional scopes are narrow. Even so, there are numerous practical applications that are morally charged and yet adequately handled by AIs.

Finally, to trust machines, they will need to be explicable. For reasons of intelligibility and accountability, AIs will need to keep logs and explanations as to why they do what they do.

Discussion Questions:

- What risks would you take with an AI and what risks would you not take? Explain.
- Try to think of ways in which explicability of algorithms or AI could be implemented or operationalised. How might an AI explain to different people?
- We expect a robot or AI system to be more fair and unbiased than a human being. Are there limitations to this expectation? Discuss.

Further Reading:

- Wendell Wallach and Colin Allen. *Moral machines*: Teaching robots right from wrong. Oxford University Press, 2008. http://www.worldcat.org/oclc/1158448911.
- Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena. Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. *Minds and Machines* 28(4):689–707, Dec 2018. ISSN 1572-8641. Doi: 10.1007/s11023-018-9482-5. URL https://doi.org/10.1007/s11023-018-9482-5.