Chapter 7
Psychological Aspects of AI

In this chapter we discuss how people relate to robots and autonomous systems from a psychological point of view. Humans tend to anthropomorphise them and form unidirectional relationships. The trust in these relationships is the basis for persuasion and manipulation that can be used for good and evil.

In this chapter we discuss psychological factors that impact the ethical design and use of AIs and robots. It is critical to understand that humans will attribute desires and feelings to machines even if the machines have no ability whatsoever to feel anything. That is, people who are unfamiliar with the internal states of machines will assume machines have similar internal states of desires and feelings as themselves. This is called anthropomorphism. Various ethical risks are associated with anthropomorphism. Robots and AIs might be able to use "big data" to persuade and manipulate humans to do things they would rather not do. Due to unidirectional emotional bonding, humans might have misplaced feelings towards machines or trust them too much. In the worst-case scenarios, "weaponised" AI could be used to exploit humans.

7.1 Problems of Anthropomorphisation

Humans interact with robots and AI systems as if they are social actors. This effect has called as the "Media Equation" (Reeves and Nass 1996). People treat robots with politeness and apply social norms and values to their interaction partner (Broadbent 2017). Through repeated interaction, humans can form friendships and even intimate relationships with machines. This anthropomorphisation is arguably hard-wired into our minds and might have an evolutionary basis (Zlotowski et al. 2015). Even if the designers and engineers did not intend the robot to exhibit social signals, users might still perceive them. The human mind is wired to detect social signals and to interpret

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_7

even the slightest behaviour as an indicator of some underlying motivation. This is true even of abstract animations. Humans can project “theory of mind” onto abstract shapes that have no minds at all (Heider and Simmel 1944). It is therefore the responsibility of the system’s creators to carefully design the physical features and social interaction the robots will have, especially if they interact with vulnerable users, such as children, older adults and people with cognitive or physical impairments.

To accomplish such good social interaction skills, AI systems need to be able to sense and represent social norms, the cultural context and the values of the people (and other agents) with which they interact (Malle et al. 2017). A robot, for example, needs to be aware that it would be inappropriate to enter a room in which a human is changing his/her underwear. Being aware of these norms and values means that the agent needs to be able to sense relevant behaviour, process its meaning and express the appropriate signals. A robot entering the bedroom, for example, might decide to knock on the door prior to entering. It then needs to hear the response, even if only non-verbal utterance, and understand its meaning. Robots might not need to be perfectly honest. As Oscar Wilde observed “The truth is rarely pure and never simple.” White lies and minor forms of dishonesty are common in human-human interaction (Feldman et al. 2002; DePaulo et al. 1996).

7.1.1 Misplaced Feelings Towards AI

Anthropomorphism may generate positive feelings towards social robots. These positive feelings can be confused with friendship. Humans have a natural tendency to assign human qualities to non-human objects. Friendships between a human and an autonomous robot can develop even when the interactions between the robot and the human are largely unidirectional with the human providing all of the emotion. A group of soldiers in Iraq, for example, held a funeral for their robot and created a medal for it (Kolb 2012). Carpenter provides an in-depth examination of human-robot interaction from the perspective of Explosive Ordinance Disposal (EOD) teams within the military (Carpenter 2016). Her work offers an glimpse of how naturally and easily people anthropomorphise robots they work with daily. Robinette et al. (2016) offered human subjects a guidance robot to assist them with quickly finding an exit during an emergency. They were told that if they did not reach the exit within the allotted 30 s then their character in the environment would perish. Those that interacted with a good guidance robot that quickly led them directly to an exit tended to name the robot and described its behaviour in heroic terms. Much research has shown that humans tend to quickly befriend robots that behave socially.

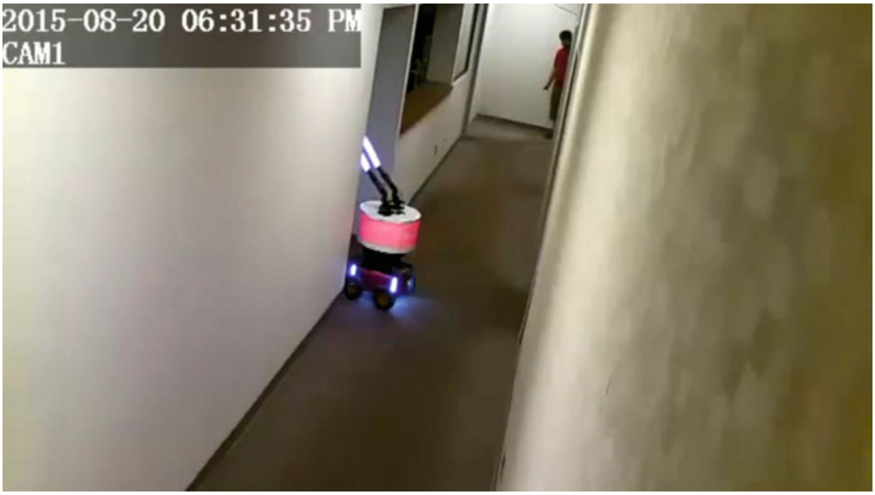

Fig. 7.1 Robot guiding people out of a building

7.1.2 Misplaced Trust in AI

Users may also trust the robot too much. Ever since the Eliza experiments of the 1960s, it has become apparent that computers and robots have a reputation of being honest. While they rarely make mistakes in their calculations, this does not mean that their decisions are smart or even meaningful. There are examples of drivers blindly following their navigation devices into even dangerous and illegal locations. Robinette et al. (2016) showed that participants followed an obviously incompetent robot in a fire evacuation scenario. It is therefore necessary for robots to be aware of the certainty of their own results and to communicate this to the users in a meaningful way (Fig. 7.1).

7.2 Persuasive AI

By socially interacting with humans for a longer period, relationships will form that can be the basis for considerable persuasive power. People are much more receptive to persuasion from friends and family compared to a car salesperson. The first experiments with robotic sales representatives showed that the robots do have sufficient persuasive power for the job (Ogawa et al. 2009). Other experiments have explored the use of robots in shopping malls (Shiomi et al. 2013; Watanabe et al. 2015). This persuasive power can be used for good or evil.

The concern is that an AI system may use, and potentially abuse, its powers. For example, it might use data, such as your Facebook profile, your driving record or your credit standing to convince a person to do something they would not normally do. The result might be that the person's autonomy is diminished or compromised when interacting with the robot. Imagine, for example, encountering the ultimate robotic car sales person who knows everything about you, can use virtually imperceptible micro expression to game you into making the purchase it prefers. The use of these "superpowers" for persuasion can limit a person's autonomy and could be ethically questionable.

Persuasion works best with friends. Friends influence us because they have intimate knowledge of our motivations, goals, and personality quirks. Moreover, psychologists have long known that when two people interact over a period of time they begin to exchange and take on each other subtle mannerisms and uses of language (Brandstetter et al. 2017). This is known as the Michelangelo phenomenon. Research has also shown that as relationships grow, each person's uncertainty about the other person reduces fostering trust. This trust is the key to a successful persuasion. Brandstetter and Bartneck (2017) showed that it only takes 10% of the members of a community to own a robot at which changes in the use of language in the whole community can take place.

More importantly, people might be unaware of the persuasive power of AI systems similar to how people were unaware of subliminal advertising in the 1950s. It is unclear who will be in control of this persuasive power. Will it be auctioned off for advertisers? Will the users be able to set their own goals, such as trying to break a bad habit? Unsophisticated people might be exploited and manipulated by large corporations with access to their psychological data. Public scrutiny and review of the operations of businesses with access to such data is essential.

7.3 Unidirectional Emotional Bonding with AI

The emotional connection between the robot or AI system and its user might be unidirectional. While humans might develop feelings of friendship and affection towards their silicon friends and these might even be able to display emotional expressions and emit signals of friendship, the agent might still be unable to experience any "authentic" phenomenological friendship or affection. The relationship is thereby unidirectional which may lead to even more loneliness (Scheutz 2014). Moreover, tireless and endlessly patient systems may accustom people to unrealistic human behaviour. In comparison, interacting with a real human being might become increasingly difficult or plain boring.

For example, already in the late 1990s, phone companies operated flirt lines. Men and women would be randomly matched on the phone and had the chance to flirt with each other. Unfortunately, more men called in than women and thus not all of the men could be matched with women. The phone companies thus hired women to fill the gap and they got paid by how long they could keep the men on the line. These professional talkers became highly trained in talking to men. Sadly, when a real woman called in, men would often not be interested in her because she lacked the conversational skill that the professional talkers had honed. While the phone company succeeded in making profit, the customers failed to achieve dates or actual relationships since the professional women would always for unforeseeable reasons be unavailable for meetings. This example illustrates the danger of AI systems that are designed to be our companion. Idealised interactions with these might become too much fun and thereby inhibit human-human interaction.

These problems could become even more intense when considering intimate relationships. An always available amorous sex robot that never tires might set unrealistic if not harmful and disrespectful expectations. It could even lead to undesirable cognitive development in adolescents, which in turn might cause problems. People might also make robotic copies of their ex-lovers and abuse them (Sparrow 2017).

Even if a robot appears to show interest, concern, and care in a person, these robots cannot truly have these emotions. Nevertheless, naive humans tend to believe that the robot does in fact have emotions as well, and a unidirectional relationship can develop. Humans tend to befriend robots even if they present only a limited veneer of social competence. Short et al. (2010) found that robots which cheated while playing the game rock, paper, scissors were viewed as more social and got more attributions of mental state compared to those that did not. People may even hold robots as morally accountable for mistakes. Experiments have shown that when a robot incorrectly assesses a person's performance in a game, preventing them from winning a prize, people hold the robot morally accountable (Kahn et al. 2012).

Perhaps surprisingly, even one's role while interacting with a robot can influence the bond that develops. Kim, Park, and Sundar asked study participants to either act as a caregiver to a robot or to receive care from a robot. Their results demonstrate that receiving care from a robot led participants to form a more positive view of the robot (Kim et al. 2013). Overall, the research clearly shows that humans tend to form bonds with robots even if their interactions with the robot are one-directional, with the person providing all of the emotion. The bond that the human then feels for the robot can influence the robot's ability to persuade the person.

Discussion Questions:

- Are there some things that you would rather discuss with an AI than a human? Create a list of general topics that might be easier to confess with an AI.
- Should robots always tell the truth, even if it results in socially awkward situations? List some situations that would be awkward.
- What is unidirectional emotional bonding? What makes it possible? Explain.

Further Reading:

- Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3-4):143–166, 2003. Doi: 10.1016/S0921-8890(02)00372-X. URL https://doi.org/10.1016/S0921-8890(02)00372-X
- Michael A Goodrich, Alan C Schultz, et al. Human-robot interaction: a survey. *Foundations and Trends in Human-Computer Interaction*, 1(3):203–275, 2008. Doi: 10.1561/1100000005. URL http://dx.doi.org/10.1561/1100000005.