Trust, expectations, and emotions in human-robot relations

Mark Coeckelbergh (Department of Philosophy, University of Twente)

If designers are asked to build human-friendly social robots, then this can be understood as involving, among other things, the ethical requirement to build social robots in such a way that humans generally trust them. But does it make sense to talk about ‘trust’ in relation to robots at all, and if so, what does it mean precisely?

Usually we talk about machines in terms of reliability. We see them as instruments we use for our purposes. As instrument they can be reliable or not. However, robots are often perceived as ‘more than machines’. For example, they may talk to the robot as if it was a human person and they may treat it as if it was a social other, an ‘artificial companion’. If the latter is the case, then it seems to make sense to talk about trust in human-robot relations.

In this paper I first present a framework for thinking about trust between humans. Based on my previous research, I present a little theory of trust, which distinguishes between two kinds of trust. A first form of trust depends on our explicit decision to trust, based on an assessment of the trustworthiness of the other. A second form of trust is implicit and rather unconscious and emotional. I speculate that the second form of trust is a kind of ‘default’ mode of trust, which is only sometimes – i.e. when something goes wrong – interrupted by more rational trust assessment.

Then I argue that in human-robot relations, at least to the extent that they take the form of social relations, the same kinds of trust-giving processes take place. Therefore, designers who want to build ‘trustworthy’ social robots must not only build reliable robots, but also create robots that meet the following normative requirements. First, the robot must perform in such a way that it becomes rational for the user to trust the robot. Second, the appearance of the robots – that is, how it looks like but also how it behaves – must be such that a form of implicit trust emerges without the need on the part of the user to assess the trustworthiness of the robot all too often.

What this means in practice, among other things, is that designers must make sure that the user forms the right expectations about what the robot can (not) do. For example, often people expect too much of the robot and when the robot does not meet these expectations, the robot is reliable but nevertheless does not receive trust from the human user. Robots and their designers can enhance trust by being more transparent about what we can expect.

It also means that there may be fundamental limits to the extent to which we can model and control trust: we might be able to model and control rational trust assessment by human and artificial agents, but it is far more difficult to model and control forms of implicit trust that depend on the appearance of the robot, since this appearance is not entirely within the control of the robot and its designer, but crucially depends on the ‘eye’ of the observer, that is on human perception, emotion, cognition, which is always unpredictable to some extent and depends on the situation.

Nevertheless, at least some progress can be made by using and setting up empirical studies of human-human and human-robot interactions. For example, robot designers can learn from how we pick up non-verbal cues about trustworthiness when looking at human and robotic facial expressions.