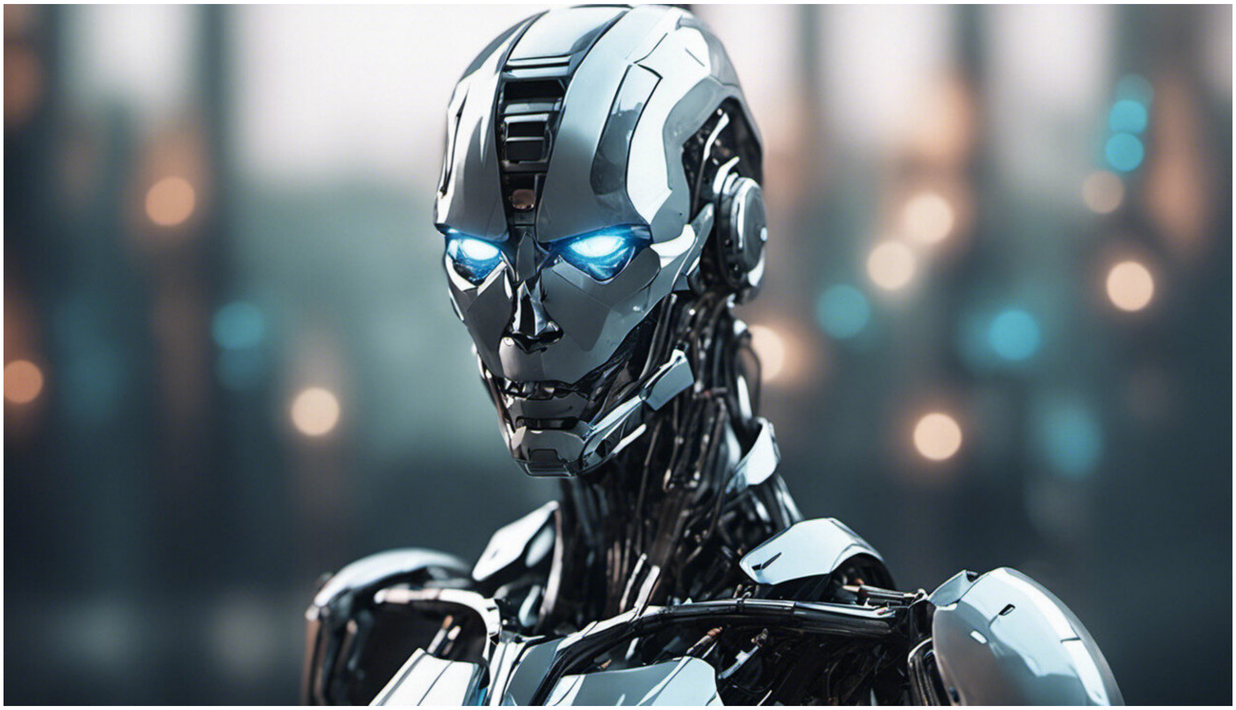
TechXplore

How to make robots that we can trust

August 29 2017, by Michael Winikoff



Credit: AI-generated image (disclaimer)

Self-driving cars, personal assistants, cleaning robots, smart homes - these are just some examples of autonomous systems.

With many such systems already in use or under development, a key question concerns trust. My central argument is that having trustworthy, well-working systems is not enough. To enable trust, the design of autonomous systems also needs to consider other requirements, including a capacity to explain decisions and to have recourse options when things go wrong.

When doing a good job is not enough

The past few years have seen dramatic advances in the deployment of autonomous systems. These are essentially software systems that make decisions and act on them, with real-world consequences. Examples include physical systems such as self-driving cars and robots, and software-only applications such as personal assistants.

However, it is not enough to engineer autonomous systems that function well. We also need to consider what additional features people need to trust such systems.

For example, consider a personal assistant. Suppose the personal assistant functions well. Would you trust it, even if it could not explain its decisions?

To make a system trustable we need to identify the key prerequisites to trust. Then, we need to ensure that the system is designed to incorporate these features.

What makes us trust?

Ideally, we would answer this question using experiments. We could ask people whether they would be willing to trust an autonomous system. And we could explore how this depends on various factors. For instance, is providing guarantees about the system's behaviour important? Is providing explanations important?

Suppose the system makes decisions that are critical to get right, for example, self-driving cars avoiding accidents. To what extent are we more cautious in trusting a system that makes such critical decisions?

These experiments have not yet been performed. The prerequisites discussed below are therefore effectively educated guesses.

Please explain

Firstly, a system should be able to explain why it made certain decisions. Explanations are especially important if the system's behaviour can be non-obvious, but still correct.

For example, imagine software that coordinates disaster relief operations by assigning tasks and locations to rescuers. Such a system may propose task allocations that appear odd to an individual rescuer, but are correct from the perspective of the overall rescue operation. Without explanations, such task allocations are unlikely to be trusted.

Providing explanations allows people to understand the systems and can support trust in unpredictable systems and unexpected decisions. These explanations need to be comprehensible and accessible, perhaps using natural language. They could be interactive, taking the form of a conversation.

If things go wrong

A second prerequisite for trust is recourse. This means having a way to be compensated, if you are adversely affected by an autonomous system. This is a necessary prerequisite because it allows us to trust a system that isn't 100% perfect. And in practice, no system is perfect.

The recourse mechanism could be legal, or a form of insurance, perhaps modelled on New Zealand's approach to accident compensation.

However, relying on a legal mechanism has problems. At least some autonomous systems will be manufactured by large multinationals. A legal mechanism could turn into a David versus Goliath situation, since it involves individuals, or resource-limited organisations, taking multinational companies to court.

More broadly, trustability also requires social structures for regulation and governance. For example, what (inter)national laws should be enacted to regulate autonomous system development and deployment? What certification should be required before a self-driving car is allowed on the road?

It has been argued that certification, and trust, require verification. Specifically, this means using mathematical techniques to provide guarantees regarding the decision making of autonomous systems. For example, guaranteeing that a car will never accelerate when it knows another car is directly ahead.

Incorporating human values

For some domains the system's decision making process should take into account relevant human values. These may include privacy, human autonomy and safety.

Imagine a system that takes care of an aged person with dementia. The elderly person wants to go for a walk. However, for safety reasons they should not be permitted to leave the house alone. Should the system allow them to leave? Prevent them from leaving? Inform someone?

Deciding how best to respond may require consideration of relevant underlying human values. Perhaps in this scenario safety overrides autonomy, but informing a human carer or relative is possible. Although the choice of who to inform may be constrained by privacy.

Making autonomous smarter

These prerequisites – explanations, recourse and humans values – are needed to build trustable autonomous systems. They need to be considered as part of the design process. This would allow appropriate functionalities to be engineered into the system.

Addressing these prerequisites requires interdisciplinary collaboration. For instance, developing appropriate explanation mechanisms requires not just computer science but human psychology. Similarly, developing software that can take into account human values requires philosophy and sociology. And questions of governance and certification involve law and ethics.

Finally, there are broader questions. Firstly, what decisions we are willing to hand over to software? Secondly, how society should prepare and respond to the multitude of consequences that will come with the deployment of automated systems.

For instance, considering the impact on employment, should society respond by introducing some form of Universal Basic Income?

This article was originally published on The Conversation. Read the original article.

Provided by The Conversation

Citation: How to make robots that we can trust (2017, August 29) retrieved 3 May 2024 from https://techxplore.com/news/2017-08-robots.html