PHYS.ORG

Finding trust and understanding in autonomous technologies

January 2 2017, by David Danks

In 2016, self-driving cars went mainstream. Uber's autonomous vehicles became ubiquitous in neighborhoods where I live in Pittsburgh, and briefly in San Francisco. The U.S. Department of Transportation issued new regulatory guidance for them. Countless papers and columns discussed how self-driving cars should solve ethical quandaries when things go wrong. And, unfortunately, 2016 also saw the first fatality involving an autonomous vehicle.

Autonomous technologies are rapidly spreading beyond the transportation sector, into health care, advanced cyberdefense and even autonomous weapons. In 2017, we'll have to decide whether we can trust these technologies. That's going to be much harder than we might expect.

Trust is complex and varied, but also a key part of our lives. We often trust technology based on predictability: I trust something if I know what it will do in a particular situation, even if I don't know why. For example, I trust my computer because I know how it will function, including when it will break down. I stop trusting if it starts to behave differently or surprisingly.

In contrast, my trust in my wife is based on understanding her beliefs, values and personality. More generally, interpersonal trust does not involve knowing exactly what the other person will do – my wife certainly surprises me sometimes! – but rather why they act as they do. And of course, we can trust someone (or something) in both ways, if we know both what they will do and why.

I have been exploring possible bases for our trust in self-driving cars and other autonomous technology from both ethical and psychological perspectives. These are devices, so predictability might seem like the key. Because of their autonomy, however, we need to consider the importance and value – and the challenge – of learning to trust them in the way we trust other human beings.

Autonomy and predictability

We want our technologies, including self-driving cars, to behave in ways we can predict and expect. Of course, these systems can be quite sensitive to the context, including other vehicles, pedestrians, weather conditions and so forth. In general, though, we might expect that a self-driving car that is repeatedly placed in the same environment should presumably behave similarly each time. But in what sense would these highly predictable cars be autonomous, rather than merely automatic?

There have been many different attempts to define autonomy, but they all have this in common: Autonomous systems can make their own (substantive) decisions and plans, and thereby can act differently than expected.

In fact, one reason to employ autonomy (as distinct from automation) is precisely that those systems can pursue unexpected and surprising, though justifiable, courses of action. For example, DeepMind's AlphaGo won the second game of its recent Go series against Lee Sedol in part because of a move that no human player would ever make, but was nonetheless the right move. But those same surprises make it difficult to establish predictability-based trust. Strong trust based solely on predictability is arguably possible only for automated or automatic systems, precisely because they are predictable (assuming the system functions normally).

Embracing surprises

Of course, other people frequently surprise us, and yet we can trust them to a remarkable degree, even giving them life-and-death power over ourselves. Soldiers trust their comrades in complex, hostile environments; a patient trusts her surgeon to excise a tumor; and in a more mundane vein, my wife trusts me to drive safely. This interpersonal trust enables us to embrace the surprises, so perhaps we could develop something like interpersonal trust in self-driving cars?

In general, interpersonal trust requires an understanding of why someone acted in a particular way, even if you can't predict the exact decision. My wife might not know exactly how I will drive, but she knows the kinds of reasoning I use when I'm driving. And it is actually relatively easy to understand why someone else does something, precisely because we all think and reason roughly similarly, though with different "raw ingredients" – our beliefs, desires and experiences.

In fact, we continually and unconsciously make inferences about other people's beliefs and desires based on their actions, in large part by assuming that they think, reason and decide roughly as we do. All of these inferences and reasoning based on our shared (human) cognition enable us to understand someone else's reasons, and thereby build interpersonal trust over time.

Thinking like people?

Autonomous technologies – self-driving cars, in particular – do not think and decide like people. There have been efforts, both past and recent, to develop computer systems that think and reason like humans. However, one consistent theme of machine learning over the past two decades has been the enormous gains made precisely by not requiring our artificial intelligence systems to operate in human-like ways. Instead, machine learning algorithms and systems such as AlphaGo have often been able to outperform human experts by focusing on specific, localized problems, and then solving them quite differently than humans do.

As a result, attempts to interpret an autonomous technology in terms of human-like beliefs and desires can go spectacularly awry. When a human driver sees a ball in the road, most of us automatically slow down significantly, to avoid hitting a child who might be chasing after it. If we are riding in an autonomous car and see a ball roll into the street, we expect the car to recognize it, and to be prepared to stop for running children. The car might, however, see only an obstacle to be avoided. If it swerves without slowing, the humans on board might be alarmed – and a kid might be in danger.

Our inferences about the "beliefs" and "desires" of a self-driving car will almost surely be erroneous in important ways, precisely because the car doesn't have any human-like beliefs or desires. We cannot develop interpersonal trust in a self-driving car simply by watching it drive, as we will not correctly infer the whys behind its actions.

Of course, society or marketplace customers could insist en masse that self-driving cars have human-like (psychological) features, precisely so we could understand and develop interpersonal trust in them. This strategy would give a whole new meaning to "human-centered design," since the systems would be designed specifically so their actions are interpretable by humans. But it would also require including novel algorithms and techniques in the self-driving car, all of which would represent a massive change from current research and development strategies for self-driving cars and other autonomous technologies.

Self-driving cars have the potential to radically reshape our transportation infrastructure in many beneficial ways, but only if we can trust them enough to actually use them. And ironically, the very feature that makes self-driving cars valuable – their flexible, autonomous decision-making across diverse situations – is exactly what makes it hard to trust them.

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

Provided by The Conversation

Citation: Finding trust and understanding in autonomous technologies (2017, January 2) retrieved 19 April 2024 from https://phys.org/news/2017-01-autonomous-technologies.html