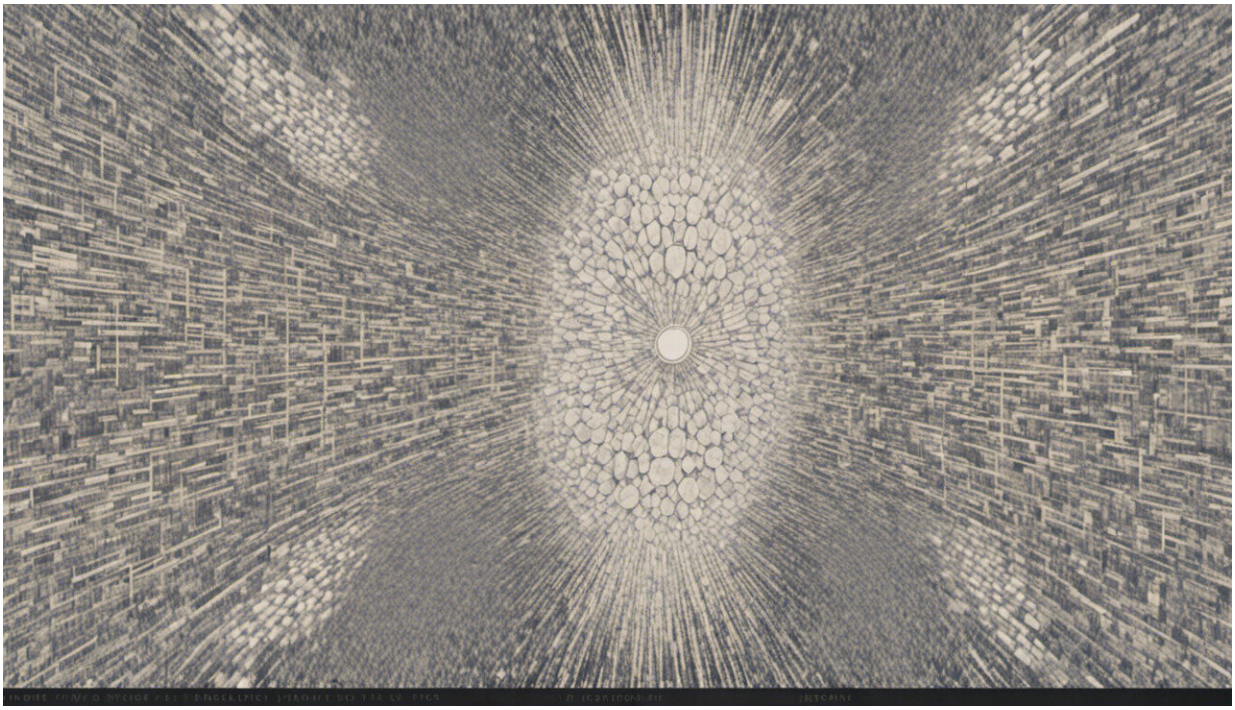
How science can help us make AI less creepy and more trustworthy

July 20 2016, by Simone Stumpf



Credit: AI-generated image (disclaimer)

Stories about racist Twitter accounts and crashing self-driving cars can make us think that artificial intelligence (AI) is a work in progress. But while these headline-grabbing mistakes reveal the frontiers of AI, versions of this technology are already invisibly embedded in many systems that we use everyday.

These everyday uses include everything from fraud detection systems that monitor credit card transactions to email filters that learn not to swamp your inbox with spam. You've probably already interacted with an AI system today without even knowing it and probably enjoyed the experience.

One increasingly common form of AI can be found in chatbots, a type of software that lets you interact with it by having a conversation. The iPhone assistant technology, Siri, is an obvious example. Microsoft's experimental Twitter account that learned how to speak from other users and ended up spouting racist phrases is another. But many websites and apps are now using chatbots to let people order services or locate specific information – without descending into bigotry.

For example, Amy is an AI assistant that schedules meetings for you via email exchanges with your contacts. Very few of these chatbots could pass themselves off completely as a human, however, so their designers need to think carefully about how people react to AI if they want their creations to be accepted. Otherwise it ends up feeling like you're talking to a really bad PA.

Teaching a machine

There are many different approaches to make these digital machines behave in an intelligent way that mimics human behaviour. But what all of them have in common is that they base what they are doing on huge amounts of data that they have gathered from their environment.

Chatbots are often "trained" by being given months of Twitter traffic as examples which is then analysed using complex statistical methods to find frequent patterns of behaviour. For example "fine, thank you" is a frequent response to a question such as "how are you?". Quite often, AI will not truly understand what it is saying, it will simply repeat what it

has seen.

Credit: AI-generated image (disclaimer)

Having a conversation with another human is actually quite complex. You need to first recognise the words in a sentence, know when it is your turn to answer, then generate your own appropriate response that relates to the point of the conversation. Several things can go wrong, from simply not knowing a word to getting the intent of the conversation wrong. Obviously, the more errors there are, the less you think the conversation is going well, and in the worst case, you might stop interacting.

We already know that people will interact differently with a human than a machine. They trust AI less, they do not engage as deeply with it, and

they will talk to it in a simpler way than with real humans. In fact, there is evidence that the more the machine tries to mimic a real human conversation, the more off-putting it is, similar to the "uncanny valley" effect that happens the more humanoid robots look.

So how can we design an AI system that is more acceptable to people? First, better and more examples of correct behaviour are needed so that it makes fewer errors. People need to start working hand-in-hand with machines to shape the behaviour of AI systems.

What also seems to matter is how much a user understands how a system works. For example, a recent study on conversational agents found that people wanted to know what the system could do, what is was doing, how it was doing it and whether it was changing due to how the user was interacting with it in the past. This point seems to apply to all kinds of AI, as transparency of an AI system seems to have a positive impact on user satisfaction.

Make it less human

Obviously, people are less likely to trust error-prone systems. But they also don't want AI to act by itself without any confirmation. For example, if you know a system often misunderstands you then you would not want it to dial a phone number without first checking it is correct. The system also needs to make clear to the user that it's a robot. It won't be like talking to another human, and that's quite ok.

We can expect to see AI systems become more accurate and more integrated into everyday life, but there will also be spectacular failures. Mostly, these systems work fine but what do we do when they don't? Since the dawn of science fiction, there have been questions about the ethics and laws of AI and how we can control it, which continue to this day. These are still open research questions that have to be answered,

along with where AI should and shouldn't be used, and who is responsible for making decisions and ultimately answerable for mistakes.

In the meantime, more and more companies are starting to integrate AI into their systems and products, with some success. Google's Nest Learning Thermostat – which memorises your schedule and changes depending on how you use it – is one obvious example but there are scores of start-ups that now leverage the power of AI to provide a personalised experience for consumers. And thanks to the rise in data science that provides the information that will teach these systems, there has never been a better time for firms to turn to the power of AI.

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

Source: The Conversation

Citation: How science can help us make AI less creepy and more trustworthy (2016, July 20) retrieved 3 May 2024 from https://phys.org/news/2016-07-science-ai-creepy-trustworthy.html