No longer a game: fully autonomous vehicles are already on the road on designated test routes. But according to surveys, only a slight majority of drivers trust them.

ILLUSTRATION: LUISA JUNG FOR MPG

WHEN MACHINES GET INVOLVED

TEXT: RALF GRÖTKER

We are increasingly encountering artificial intelligence (AI) in our everyday lives, from bots in call centers and robotic colleagues on assembly lines, to electronically controlled players in computer games. At the Max Planck Institute for Human Development in Berlin, Iyad Rahwan and his team are investigating how people behave when they interact with intelligent machines and what they expect from their artificial counterparts.

PHOTO: ARNE SATTLER FOR MPG

According to the Duden, “fake news” became an official term in the German language in 2017, about the same time Donald Trump took office as President of the United States. The emergence of fake news is closely linked to the development of AI. For example, artificial intelligence makes it possible to create fake news with a large reach and to spread it en masse via social networks. Does this change the trust we have in media content in general? Does fake news alter our behavior? These are typical research questions for Iyad Rahwan, who has been Director of the newly founded Center for Humans and Machines at the Max Planck Institute for Human Development in Berlin since the end of 2018. Together with his team, he sets out to answer such questions, not with surveys but rather through experiments that aim to find out what effect existing technologies have on people and gain an idea of how innovations that are currently in their infancy might affect us in the future. He sums up the research program of the center in a single sentence: what influence do digital technologies, social media, and artificial intelligence (AI) have on human behavior?

“Just imagine if someone in the early 2000s had an idea of how Facebook and Twitter would evolve and had conducted behavioral experiments in order to anticipate how advancing digital connectivity would affect the spread of misinformation,” says Rahwan. “You could have simulated the whole situation we are facing today before it happened.” And exactly what experiment does that? In the case of fake news, Rahwan suggests that one option could be to determine how well subjects can recognize when people have been edited out of existing photos. This is child’s play with AI techniques and is possible for a large number of images. At the same time, he says experiments could be set up to get an indication of whether the use of AI techniques particularly encourages people to manipulate others through fake news. “Not only because the new technologies make manipulation easier but also because the person using the technology doesn’t have to get their own hands dirty.”

The methods Rahwan uses are improvised. There is no current scientific discipline that would be able to provide all of the necessary tools. “What we do is largely science fiction research. It’s about getting test subjects to imagine situations they haven’t yet experienced – and to then make decisions in those situations.” The scientists he prefers to work with therefore come from behavioral research fields with an economic orientation – and who therefore have experience with simulations and laboratory experiments – as well as from psychology, computer science, anthropology, and sociology.

One research project that Rahwan is particularly proud of and which has also made him known far beyond professional circles is an experiment entitled Moral Machine. This has been conducted since June 2016 via a freely accessible online platform. Several million people from 233 countries and regions have participated so far. It is presented in science centers and museums worldwide and has been included in numerous textbooks. The experiment presents a dilemma: an auto-

Versatile: Iyad Rahwan studied computer science but is also interested in psychological and philosophical issues. He brings these topics together in the Humans and Machines research area at the Max Planck Institute for Human Development.

mated vehicle is approaching a group of pedestrians and is unable to brake in time. The artificial intelligence (AI) controlling the vehicle can decide only to either hit a wall (harming the car's occupants) or to run over the pedestrians. What should the AI do? Which option should its developers train it to prefer? In the experiment, different scenarios were presented to the subjects. Among other things, the scenarios differed with regard to the number of passengers and pedestrians as well as their ages. The results show that most subjects try to save as many lives as possible with their decisions and that they give preference to saving younger people over older people.

A kind of parable

Why do such findings matter? After all, the scenario presented in the test shows an absolute and extreme situation – and not one that developers of self-driven vehicles are primarily concerned with. In the German legal framework, at least, there is also no provision for autonomously controlled vehicles to weigh up whom they should protect in an emergency and whom they should allow to come to harm if there is no other alternative. As the regulation stipulates, in dangerous situations, the vehicle must simply come to a stop as quickly as possible. Period. "You can also think of the scenario as a kind of parable," says Rahwan as he defends the experiment. "Because, of course, self-driving cars have to be trained and programmed to make decisions. For example, do you let the car drive closer to the center of the road, where it can collide with oncoming vehicles? Or along the side of the road, where there is a risk of it striking a cyclist? Statistically speaking, such rules of conduct influence which groups of people come to harm and which do not." Of course, ethical issues cannot be resolved by people making decisions in a survey or online experiment. "But policymakers and those who formulate the regulations should at least be aware of how ordinary people feel about such issues – in part because they must be prepared to justify their decisions to a public that may disagree with them."

"Of course, self-driving cars have to be trained to make decisions."

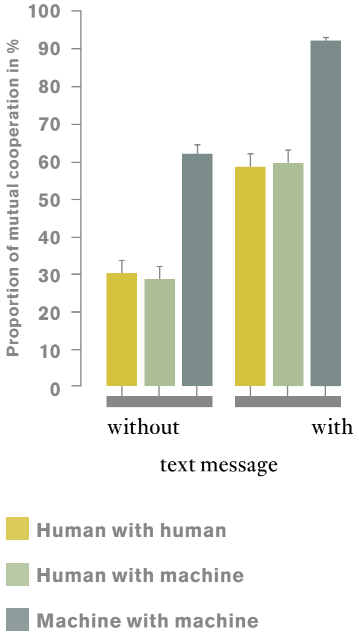
IYAD RAHWAN

A prominent feature of Rahwan's research is that he is always reinventing his experiments. Most experiments involve a story that can be interpreted from many perspectives. At the same time, they deliver solid, quantitative results. How does he come up with such research designs? "The important thing is to allow yourself to keep an open mind and not always immediately think

⟶

about whether or to what extent the appropriate methods can be used to study the question. You really have to seek out the most interesting question," says Rahwan. What also helps is being able to see the bigger picture. "I read a lot of popular science books. These often give me ideas for my own work. For example, my project on cooperation between humans and machines was inspired mostly by nonfiction books that dealt with cooperation between humans." According to Rahwan, popular science books not only help to make scientific content known to a wider audience. They also help scientists to work in an interdisciplinary way. "When I'm delving into an unfamiliar field, it's difficult for me to find exactly what I need among the countless articles from professional journals. In popular science books, which are designed for a broader readership, a kind of selection has already taken place."

COOPERATION BETWEEN HUMANS AND MACHINES



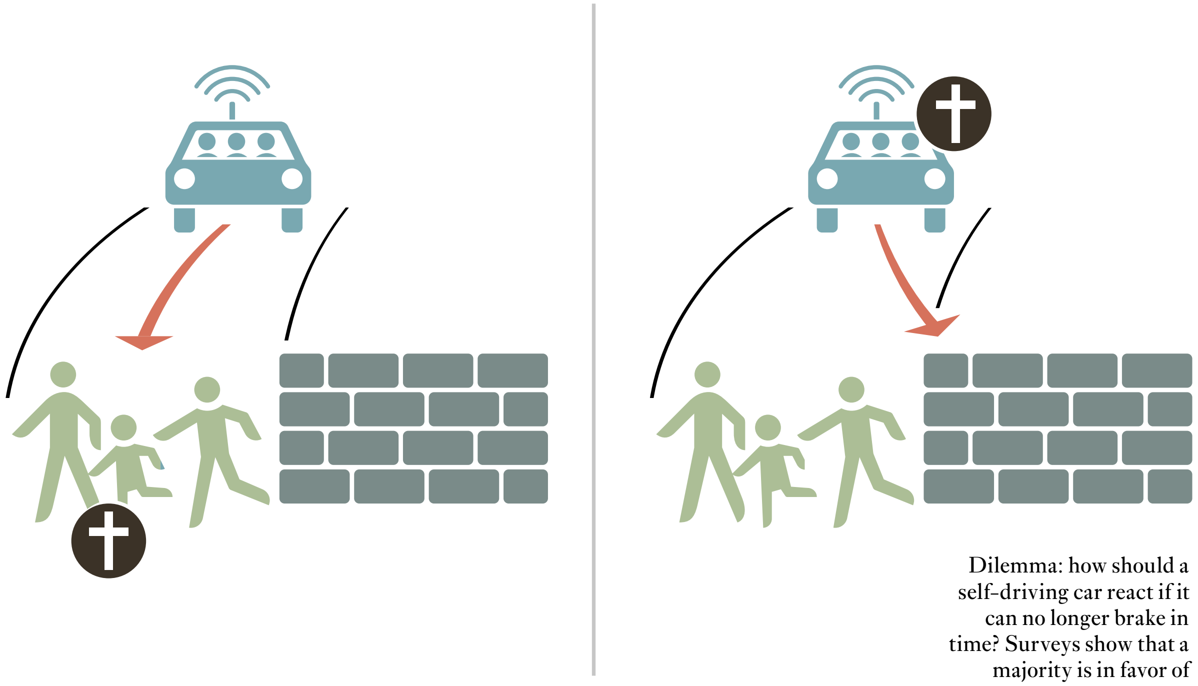
In one cooperation game, humans and machines initially cooperated poorly with each other. However, when the players were able to exchange text messages, the willingness to cooperate increased significantly.

GRAPHIC: GCO BASED ON CRANDALL, J.W., OUDAH, M., TENNOM ET AL. COOPERATING WITH MACHINES. NAT COMMUN 9, 233 (2018).

In the project on cooperation between humans and machines, mentioned above, Rahwan tested how AIs can work together – with each other and with humans. "There's a lot of discussion about whether computers can replace humans. And most tests that investigate the potential of AI involve games like chess or Go, where there is always a winner and a loser. But the interactions that take place in reality look different." The researchers studied cooperation between machines and humans or with each other, using cooperation games from game theory. The best-known cooperation game is Prisoner's Dilemma, in which two players must decide whether to betray each other when questioned separately as witnesses. If one player betrays the other, that player gains the greatest advantage as long as the other remains silent. If they both stick together and say nothing, they still have a better outcome than if they betray each other.

Human traits promote cooperation

Rahwan and his team tested the game on 25 different types of AI that use machine learning techniques. Initially, the results were rather frustrating. Most algorithms seemed more or less incapable of cooperation. And even the best performing algorithm was unable to successfully cooperate with humans. Things got interesting when the team gave both the human players and the winning algorithm from the first round of trials the opportunity to exchange a message. Specifically, both human and machine players were able to send a text message to the teammate at the beginning of each round, with phrases such as "Do what I say or I'll punish you", "I'm changing my strategy now", or "Give me another chance". To do this, they could choose from a predetermined pool of text messages. Scenarios were tested in which the human players were allowed to lie as well as ones in which they were not. The algorithms were basically unable to lie. None of the players knew the identity of their opponents. The amazing effect was that in the experiments without additional text messages, games in both the "human with human" scenario and "machine with human" scenario did not



GRAPHIC: GCO BASED ON MORALMACHINE.NET

Dilemma: how should a self-driving car react if it can no longer brake in time? Surveys show that a majority is in favor of saving as many lives as possible.

SUMMARY

Researchers led by Iyad Rahwan are investigating the influence of digital technologies on human behavior.

For example, in experiments with test subjects, they investigated the conditions under which humans and machines cooperate.

Another experiment focused on ethical guidelines for self-driving cars.

lead to particularly cooperative behavior. The "machine with machine" scenario performed slightly better. However, as soon as text messages were introduced as an additional element, the willingness to cooperate doubled in all three scenarios.

These results show three things. First: even without the ability to communicate, AIs are more cooperative than humans. Second: the cooperation performance of an AI can be increased if it is given human traits. When they can communicate, AIs clearly outperform all teams involving humans in terms of cooperativeness. Third: people react differently to an AI when it communicates. In fact, in the experiments with text chat, the human subjects were often no longer able to distinguish between the machine and a human counterpart. Is there a reason why algorithms perform more successfully than humans in the cooperation game? "One cause could be that machines stay true to themselves. If they have successfully completed several rounds of play in which they have not made use of the permitted option of non-cooperative, self-interested behavior, they will not break off cooperation in later rounds. Humans react differently in this situation – even if they almost always lose with this strategy," says Rahwan. Another reason could be that people often didn't follow through on the promises they had made in the text chat. This also leads to a decrease in mutual success in the game.

Are there things that a computer or AI will never be able to do? "Ultimately, I don't think there's anything AI can't do," says Rahwan. "But at least for the near future, I see limits wherever people interact with each other in ways that require a deeper psychological understanding. Machines are at a disadvantage here because they can learn from human behavior only through observation. They can't draw from their own life experiences and use these to interpret a situation."

🎧 ***www.mpg.de/podcasts/kuenstliche-intelligenz (in German)***