By Invitation | Artificial intelligence

We need to focus more on the social effects of AI, says Nicholas Christakis

The sociologist's experiments suggest it will change how humans treat each other

IMAGE: DAN WILLIAMS

Dec 15th 2023

FEW PEOPLE would tolerate a virtual assistant if they had to plead obsequiously each time, “Excuse me, Alexa, if it’s not too much trouble, could you kindly tell me what the weather will be today.” Instead, these devices are designed to answer brusque commands: “Alexa: weather!” And we expect them to obediently respond.

Which is fine until we bring them into a home with impressionable young children, who may quickly learn that this is a normal way to talk to other people—that is, rudely. This points to a potentially far-reaching problem with artificial intelligence (AI). When it comes to how AI will affect social interaction, most people are focused on the relationship between humans and AI. Not enough attention is being paid to how humans will treat each other in the presence of AI.

Unlike AI used for technical challenges, such as processing medical images, certain types of AI are designed to act in more human ways, like providing psychotherapy. These technologies will induce “social spillovers”—influencing how people react to and learn from the behaviour of other people. And these spillovers might affect humans well beyond those involved in the original interaction.

People will increasingly have AI-enabled "co-bots" on their phones that get to know them and help them relate to other people. But some users of dating apps, for instance, have found that they enjoy flirting with a virtual partner more than going on an actual date. This changes the sorts of people available in the real, human dating pool, in addition to reshaping interpersonal communications.

Although chatbot conversation partners and other types of "smart" AI powered by large language models (LLMs) may seem the most consequential for human behaviour, even small intrusions into our social lives by simpler AI can have profound spillover effects, for good or ill.

In one experiment, we placed 4,000 people into 230 online groups. Each group was then divided into several clusters, each with just a few people. The members of these clusters had to co-operate with each other on picking colours. If they found a "solution"—with each individual choosing a different colour than their immediate neighbours—the group as a whole was said to have succeeded and everyone got some money.

To some of these groups, however, we surreptitiously added bots that the members perceived to be other humans—and manipulated their responses. We found that having the bots occasionally make "erroneous" moves that increased rather than decreased the colour conflicts with their immediate neighbours was actually helpful to the group as a whole, fostering greater flexibility. People came to realise that just solving the challenge with respect to themselves and their immediate neighbours was not necessarily best for their group as a whole. Making a counterintuitive move that seemingly decreased local consensus unlocked a group-wide solution. The AI was able to help the people to help themselves.

In another experiment, we gave 1,024 subjects in 64 groups the challenge of producing so-called public goods—items that people work together to fashion and that are of mutual benefit, like a lighthouse. The idea is that if everyone pitches in, everyone will end up benefiting more than they contributed. But, of course, the temptation is to let *others* work to tend the commons.

At the beginning, over 60% of people acted altruistically and helped out. But we found that by adding just a few bots (which the players again perceived to be other humans) that behaved in a free-riding way, we could drive the group of people to behave selfishly so that, eventually, they stopped co-operating altogether. The bots could convert a group of people who were otherwise generous into a group of jerks.

But the opposite was also true. We could use bots to enhance human could use bots to enhance human co-operation. Giving people co-operative (artificial) partners caused them to be kinder than they would normally be when dealing with other people.

Other experiments show that when people delegate decision-making to AI agents—something they are increasingly likely to do, from having LLMs draft emails to tasking drones with military targeting—it can obscure moral responsibility and encourage unethical interactions with other people.

A group at the Max Planck Institute for Human Development led by Iyad Rahwan has done experiments that involved giving subjects AI assistants. People had to roll dice and report the outcome. Around 5% of the participants were dishonest when doing the task by themselves. That number rose to 40% when subjects could delegate the task of being dishonest to another human, and to 50% if they could delegate it to a machine. But the number rose to a whopping 88% if they could delegate the task to an AI agent that could decide to cheat on their behalf.

If undermining honesty as people interact is not worrying enough, there are fears that AI could undermine physical safety, too. In just-published experiments led by Hirokazu Shirado at Carnegie Mellon University, we found that even very simple forms of AI assistance for drivers, such as auto-steering or auto-braking, eroded social norms of reciprocity on the road. Allowing humans to delegate whether to swerve away from an oncoming car in repeated games of chicken resulted in people subsequently being less likely to take turns in giving way, thereby increasing the frequency of crashes when they drove without AI assistance.

These effects of AI suggest that it could have a big impact on the social norms that have evolved over millennia, shaping how we treat each other in all manner of everyday interactions. Governments cannot afford to ignore the risks. At a minimum. thev should evaluate more closely whether AI systems are aligned with human social interests and they should provide for more safety testing. As the Bletchley Declaration signed at the recent AI-safety summit in Britain made clear, innovation must go hand in hand with attention to mitigating risks. After all, we cannot ask AI to regulate itself, even politely. ■

Nicholas A. Christakis is the director of the Human Nature Lab at Yale University.