



ChatGPT is posing challenges for educators as well as lawyers and artists on issues from plagiarism to ownership rights. But at the heart of these problems is one about ethics and responsibility – how can ChatGPT and other new technologies be used to enhance society, rather than to endanger or undermine it? PHOTO: AFP

The Petrov dilemma: Moral responsibility in the age of ChatGPT

The way a Soviet officer dealt with potential nuclear armageddon holds an important lesson in how we deal with the myriad challenges thrown up by ChatGPT and others of its kind

Julian Savulescu

In 1983, Stanislav Petrov "saved the world". At the height of the Cold War, Lieutenant-Colonel Petrov was the duty officer at a Russian nuclear missile facility when the early warning system flashed red with a heart-stopping alert: The United States had launched a missile strike and five Minuteman missiles were streaking towards the Soviet Union.

According to protocol, Lt-Col Petrov should have reported the imminent strike. If he had, Soviet nuclear doctrine called for full nuclear retaliation. But after a nerve-wracking five minutes, he disobeyed orders.

He had his reasons. There was no corroborating evidence from ground-based radar and the missile defence system was new. The number of American missiles was also, to him, suspiciously low for an all-out US first strike.

If he had acted on the information given by the machine, the result would have been catastrophic – an estimated death toll of nearly 290 million in the US and Soviet Union and millions more worldwide – from starvation and other spillover effects of nuclear war.

In this instance, faulty new satellite technology set off a false alarm. As a human decision-maker, Lt-Col Petrov's appreciation of the gravity of nuclear war influenced – and enhanced – his judgment. It overrode standing orders and the threat of execution for treason for disobeying them.

In contrast to nuclear missile early-warning systems, ChatGPT does not come laden with doomsday possibilities. At a basic level it is a piece of machine intelligence that gathers information from billions of human conversations and

produces text in response to prompts.

And yet in its own way ChatGPT's abilities and how we deal with it are also potentially world-changing, even if not on the scale of nuclear armageddon. The Petrov dilemma is pertinent too as we confront new challenges posed by artificial intelligence (AI).

Consider the world before the arrival of the Internet. It was a very different one with none of the benefits and ills that successive waves of technological breakthroughs and innovations have brought to billions of people around the world.

Technology has made possible swift access to information, public health, banking and other services. The flip side is we are struggling with its dark side: fake news, deep fakes, online scams and revenge pornography.

ANOTHER IPHONE MOMENT

And now we have ChatGPT, the arrival of which has been described as another iPhone moment. As analyst Rowan Curran describes it, the launch of Apple's iPhone in 2007 and the App Store a year later ushered in a period of historic technological change by putting an entire computing experience in our pockets. The ease of use also opened up to everyone vast new opportunities in education, work and play.

ChatGPT appears to put us on the cusp of yet another game-changing moment in our relationship with machines. Our increasing reliance on them has saved lives, averted air disasters, enhanced medical diagnosis and, in time to come, perhaps reduce car accidents.

But like Lt-Col Petrov, we still need to interrogate what the machines tell us. There is still a place for human decision-making when it comes to moral choices in a world of machines.

HOW SHOULD WE DEAL WITH AI?

ChatGPT is posing challenges for educators as well as lawyers and artists on issues from plagiarism to ownership rights. These are important matters and need to be addressed. But in a way they are secondary problems. At the heart of these problems is one about ethics and responsibility – how can ChatGPT and other new technologies be used to enhance society, rather than to endanger or undermine it?

One approach is to view

challenges through a set of three guiding principles on technological interaction: Think First, Take Responsibility, Act Ethically.

THINK FIRST

Lt-Col Petrov did not just accept information he was presented with from an expert system. He actively interrogated it. He asked whether it was consistent with other beliefs he held. When presented with empirical claims about the world by machines, we should ask: are they true? Are they consistent with other beliefs? How confident am I that they are true? Should I gather more information?

In life, trust is necessary as it is not possible to test every claim presented to us. Yet whom and what to trust has become more difficult with the proliferation of dodgy information, a problem that is likely to worsen with ChatGPT and its like.

To counter this, we need to do more to teach the basics of critical thinking to children, equipping them with the skills to question the basis of assertions and think logically and rationally.

Sometimes we are not presented with claims about the way the world is but normative claims: claims about how it should be. That is, claims about ethics, about right and wrong, good and bad.

Our humanity resides in our capacity to make these ethical decisions. We should ask: what are the ethical reasons to accept this normative claim? To arrive at the best ethical judgment requires that we bring together our ethical theories, principles and concepts to a particular issue, along with the best evidence about the facts. We should aim for maximum consistency across our normative beliefs about the way things should be.

There are many calls for AI to be explainable. What is even

more important than explainability is justifiability. Justifiability is being able to provide reasons based on values we hold to perform an action.

TAKE RESPONSIBILITY

We are not responsible for things we can't avoid or foresee. But if an outcome is foreseeable and avoidable, we are responsible for it. And we are equally responsible for outcomes of our omissions when we choose to do nothing. Choosing to accept the information presented to us, or to follow a directive, is a choice for which we are responsible.

Stanislav Petrov took responsibility: he chose not to act. When presented with an empirical or normative claim by an AI such as ChatGPT, we should ask: do I have good reason to act on this claim? If a piece of technology, like a calculator, is highly reliable, then the answer is yes. But for novel technology, the issue is much murkier.

Blame is a function of the moral responsibility of a person and the degree of harm that results. Praise results from responsibility and benefit. The degree to which someone is responsible also depends on the effort or commitment they exert to bring about a benefit or avoid a harm.

Much of the debate around plagiarism and authorial credit and ChatGPT assumes people will cheat: exert no effort and take credit for value created. One major scientific journal, Science, has banned any output from ChatGPT.

But ChatGPT can involve human interaction. It represents data of a certain quality. Co-creativity is possible provided that significant effort is expended in the production, refinement or evaluation of the outputs of AI. If a person "mixes their labour" with ChatGPT, they can be deserving of praise, just as computing power has enhanced

our ability to produce new mathematical and scientific research, not undermined its validity or value.

ACT ETHICALLY

What is to act ethically has been debated for thousands of years. But at its heart is a simple answer: add value to the lives of others. For Lt-Col Petrov, doing nothing was the hardest choice: "All I had to do was to reach for the phone; to raise the direct line to our top commanders – but I couldn't move. I felt like I was sitting on a hot frying pan."

ChatGPT is an extremely powerful technology with enormous potential for human benefit. It can distil human knowledge and wisdom, quickly create factual and normative content, enable communication between speakers of different languages, and ultimately enhance or even surpass human creativity.

As with any powerful technology, there are significant risks. At present, there are numerous mistakes and false claims, but these presumably will be ironed out. Longer term, risks include threats to privacy; large-scale perpetuation of biases; plagiarism; displacement of human workers and creativity; deception and malicious use.

Acting ethically requires realising the value of technology while ensuring the risks are reasonable. Institutional or structural changes may be necessary, for example, labelling and creation of watermarks to prevent plagiarism.

TO BE HUMAN IS TO MAKE YOUR OWN ETHICAL DECISIONS

One ethical question that confronts regulators and industry is: should a driverless car swerve and sacrifice an innocent bystander (pedestrian or cyclist) to save a greater number of lives in the vehicle? I asked ChatGPT whether it is right to kill one to save five. Here is the answer:

"Whether or not it is right to kill one person to save five people is a complex ethical question. Some people may argue that taking one life to save five is the more logical option, and is therefore justified. However, it is also important to consider the ethical principle of respect for autonomy."

"This principle states that all individuals have an inherent right to make autonomous and informed decisions about their lives, including their own death."

Taking this into consideration, it may not be morally right to force someone to accept the sacrifice of their life in order to save the lives of five other people. Ultimately, the answer to this question is subjective and depends on one's own ethical and moral beliefs."

But as driverless cars are rolled out, we won't have the luxury of remaining on the fence, or relying on drivers' reflexes to get us out of a decision. We will have to make moral choices and give reasons, and apply them consistently to programming the vehicles. This requires deep – and human – ethical thinking and expertise.

AI is a tool but it is not conscious and it lacks the capacity to make moral decisions.

To be human is to make your own ethical decisions. Stanislav Petrov took responsibility in perhaps the grandest way possible. But each of us must take responsibility when we are presented with information and advice: are there good reasons to accept it? What should I do? This doesn't change just because a smart new piece of technology becomes available. The same applies to chat. ChatGPT may be able to produce Shakespearean-like sonnets one day but it will be up to us to decide if they are good or bad.

The challenge presented by ChatGPT is not unique – it is a problem raised by the Internet, social media and modern life in general.

With power, comes responsibility. The decision to do nothing accrues responsibility when action is a possibility. We are responsible for the consequences if we choose to do nothing, or simply follow orders. Sometimes we must follow rules, authorities, law or orders: but we should always ask, are they right? AI, such as ChatGPT, can provide us with means to achieve our ends. But it should never dictate to us what our ends should be.

Only humans can make decisions about good and bad, right and wrong. The biggest threat of AI is that it results in dehumanisation as we blithely accept its output without taking a first person stance and evaluating the justifiability of its output. We must be active participants engaging with technology, not passive consumers.

* Julian Savulescu is Chen Su Lan Centennial Professor of Medical Ethics, and director of the Centre for Biomedical Ethics, Yong Loo Lin School of Medicine, National University of Singapore.