COMMENTARY

Killer Robots and Deepfakes: Activists and artificial intelligence

Brian Martin

If you're holding a rally, what do you do when confronted by automated police agents demanding that you leave? Can you negotiate with them? What if they start firing their weapons?

Another challenge: an opponent has produced a realistic video showing your group's most prominent figure in a compromising position. It's being widely circulated on social media. What can you do?

These are just two of the challenges potentially posed by artificial intelligence (AI), one of the most important technological developments of our time. Computing systems can be developed to undertake complex tasks automatically and 'intelligently', such as scanning a photograph and comparing it to a database of photos, far faster and often more accurately than any human.

To give a sense of the potential challenges posed by AI, two areas are addressed here: killer robots and deepfakes. The more general challenge is for activists to be better informed and prepared for these and other AI-related developments.

Killer Robots

Militaries have been greatly interested in battlefield applications of AI. The result is what has been called the automated battlefield.

AI enables the development of aeroplanes, tanks and robots that can be controlled by soldiers at a distance and, if desired, make automated decisions about battlefield operations. Already the Israeli military has deployed the Harpy, a drone that hovers over areas and, when it recognises signals indicating a radar installation, automatically launches a weapon to destroy it. Technologists have produced swarms of small drones that communicate with each other to collectively evade missiles and try to destroy a target.

Paul Scharre's 2018 book *Army of None* is an informative tour of the technologies and issues involved with autonomous weapons. Scharre provides a history of weapons development, including WWII efforts by the Nazis. The basic story is that rapid strides in artificial intelligence are making possible weapons that are semi-autonomous or autonomous. Especially useful is Scharre's analysis of what constitutes an autonomous weapon, with illustrations about the roles that humans play in the OODA (observe, orient, decide, act) loop. The question is about where in the process human decision-making plays a role: in selecting targets, authorising strikes and/or designing algorithms.

A key consideration when developing and deploying autonomous weapons is avoiding too many civilian casualties. In controlled environments, where the target is well identified, weapons can be given more autonomy. In a cluttered and fast-changing environment, identifying a target is far more challenging. According to Scharre, US air and naval forces seem opposed to autonomous weapons, perhaps in part because they want to keep their personnel involved. An important lesson Scharre raises is that many people in the military are acutely aware that there is a potential for blowback from actions by autonomous weapons, especially if there are civilian casualties.

Deane Baker (2022) in his book *Should We Ban Killer Robots?* addresses ethical concerns raised about LAWS, which stands for lethal autonomous weapons systems. Critics, for example the group Stop Killer Robots, want these weapons banned or regulated. One of their arguments is that humans need to be in control of decisions about using lethal force in wars. Baker, countering this, points out that soldiers often have inadequate information or time to make careful decisions. For example, jet pilots may launch a missile at a target with only limited information and less than a second to make a decision. Why, he asks, are LAWS fundamentally any different? Indeed, an AI-based weapons system may be able to do much better than a human in making split-second choices about targeting and deployment.

Then there are land mines, now banned but still widely used. Land mines are autonomous in that they explode without a human intervening. With autonomous weapons, key decisions are in constructing and programming, and not necessarily so much in battlefield operation.

Baker uses a just-war framework for his analysis. Many in the peace movement are opposed to all arms manufacturing and military deployment. They would prefer to get rid of all weapons. Baker's point is that if you're going to accept many of the weapons currently in use, then LAWS are not different in any substantive sense.

There is some research on people's feelings about autonomous weapons systems, showing that they sheet home responsibility to those who manufacture and deploy the systems (Rosendorf et al. 2022: 177). This means that if LAWS do anything that seems unfair, the public reaction may be quite negative.

Robots vs Protesters?

What are the implications of autonomous weapons for nonviolent campaigners? It's useful to consider different forms of action. There are no obvious applications of weapons to deal with strikes and boycotts. The most likely area of application is policing public protests. It's possible to imagine police robots assigned to monitoring crowds or defending buildings from incursions. The next step would be that these police robots are armed and able to make quick decisions based on AI capabilities. Should a police robot see a threat, for example someone aiming a rifle, it might be programmed to counter the threat. Based on a misperception, the robot might fire on protesters.

How would this change the dynamics of a public protest? It is already the case that police are armed, with the ability to shoot protesters, and sometimes they do. There are several famous instances in which police opened fire on unarmed protesters, for example the 1960 massacre in which South African police shot protesters in the town of Sharpeville, killing perhaps one hundred of them (Frankel 2001: 150). Would it make any difference if, instead of human police, the killings in some future Sharpeville were carried out by robots using AI?

At Sharpeville, the police tried to hide evidence of their actions, including that they had shot many of the protesters in the back while they were fleeing and that they had used so-called dum-dum bullets, banned at the time. However, journalists were present and their stories and photos became front-page newspaper stories internationally.

If the same scenario eventuated but with killings by police robots, the key to generating outrage would be publicity: credible stories showing what really happened. There would be a blame game, with government and police leaders being blamed and most likely trying to avoid responsibility for killings.

When activists confront robots in the streets, some protesters might think it is satisfying to attack the robots, given that they have no feelings. Yet this might be unwise because it could put protesters in a bad light, as being violent and aggressive, therefore justifying heavy-handed policing. It is useful to think of protester actions as messages to observers, whether the actions relate to humans or robots.

Robots could be operated remotely or programmed to operate autonomously. To the extent that they operate autonomously, they are likely to be programmed to respond according to protester actions, just like humans might. Accordingly, activists might be best advised to treat robots just as they would treat human police, and even to seek to win them over. At the very least, protesters can try out different responses and learn from the interactions. One thing will remain much the same: the likely reactions of observers watching the engagements.

In summary, although it might seem that ‘killer robots’ pose a new and special threat to nonviolent activists, in practice the dynamics of interacting with robot police and soldiers may not be all that different from interacting with human ones. The rise of automated weapons systems makes it even more obvious that having no weapons is a good way to discourage being attacked.

Deepfakes

Imagine that someone could make a video showing you doing something horrible, something that would make others think less of you. For a nonviolent activist, this might show them slapping someone in the face, kicking a puppy or aiming a rifle at police, thereby discrediting their commitment to nonviolence. It’s now possible to make fake videos like these, so convincing that they are almost impossible for ordinary viewers to tell from the real thing.

Fake evidence has a long history and has often been used against activists. However, creating convincing fakes wasn’t all that easy. Visual evidence is especially important because most people tend to believe what they see. Photos — the old analogue ones taken with film — can be staged, but it requires a lot of work to make them convincing. Once taken, old photos can be altered, but it’s a tedious process. Stalin tried to reconstruct history by having some individuals, ones he had purged, painstakingly removed from photos.

Digital photos are much easier to manipulate. You just go to a site like FaceApp and alter your image, for example removing blemishes and making yourself older or younger. That’s done with AI.

Next are videos, which can be considered a series of photos that give the impression of motion. One type of manipulated video is called a shallowfake. It involves splicing together video segments to give the impression of continuity or connection. A widely shared shallowfake shows the actor Dwayne Johnson, known as The Rock, singing a song, cutting back and forth to Hillary Clinton, who is apparently listening (Skitz4twenty 2016). This involved splicing together segments from two separate videos of The Rock and Clinton. A superficial inspection reveals that Clinton was not there at the time. Yet many viewers accepted the shallowfake as showing an actual interaction. They were so prepared to think badly of Clinton that they ignored the contrary evidence before their eyes. Later, when The Rock supported Clinton in the 2016 presidential election, these gullible viewers were confused (Grothaus 2021: 28-40, 52).

In contrast to shallowfakes, deepfakes involve reconstructing digital images. Transforming the large number of digital images in a video requires far more computing power than for a single photo. But it can be done, and is becoming easier.

One technique uses what are called Generative Adversarial Networks. One AI program takes a video of you talking and tries to produce a fake video of you saying something else. Another AI program, the adversary, tries to figure out which video — the actual one or the fake — is real. Then the first program tries to do better, and so on until the fake is totally convincing.

There are many possible uses and implications of deepfakes (Karnouskos 2020). So far, the most common use of deepfake technology is for pornography. The face of a famous female actor, such as Scarlett Johansson, is used to replace the face of a female porn star, the result being a fake porn video of Johansson. It is now possible to produce videos like this with readily available technology that requires little editing skill. Needless to say, such videos are made without seeking permission from either the celebrity or the porn star.

Political uses are also troubling. One video, used to show the capabilities of the technology, shows former US president Barack Obama swearing (Monkeypaw Productions 2018). His lips and his speech are digitally created based on AI trained on videos of Obama, of which there are many. In the Ukraine war, Russians circulated a deepfake video of the Ukrainian president calling for his troops to surrender.

The technology has advanced so much that as little as a minute of video of someone talking can be sufficient to enable the creation of a realistic deepfake. This means that anyone who has been interviewed online or who has posted a video of themselves is vulnerable.

As commentators have noted, high-profile targets of deepfakes, such as celebrities and politicians, usually have access to ways to discredit the fakery, through various media outlets. On the other hand, individuals who are less prominent have fewer means to resist. For example, when a former partner seeks revenge by creating a deepfake video and circulating it to friends and co-workers, the damage can be immense and the opportunities for replying may be limited.

What about activists? Opponents may try to use deepfakes to discredit an organisation or a movement, cause distress to individuals, or encourage internal disputes. Imagine a deepfake video designed to suggest hypocrisy, showing climate activists flying private aeroplanes, nonviolent activists throwing bricks at police, or animal activists shooting elephants. Deepfake porn could be used to humiliate individual activists. Deepfakes might show bribery, illegal drug use or other crimes.

Deepfakes as methods of attack are one thing. As these become more common, there is another implication: audiences potentially may become more sceptical of photos and videos, so a real image is dismissed as possibly fake. The murder of George Floyd by police officer Derek Chauvin was recorded on a phone; this visual documentation was an important part of what triggered massive outrage. Would it have done so if audiences were exhausted by having to decide which videos are real and which ones fake? The loss of trust in recorded evidence could be the biggest impact of deepfakes (Fallis 2021).

What to Do?

Activists, like everyone, will need to understand the uses of deepfakes, for good and bad. To prepare for the wide range of possibilities, it would be worthwhile for some members to inform themselves sufficiently so they can lead discussions of the implications of deepfakes for planning campaigns, gaining support and defending against attack. More deeply, the rise of deepfake technology provides motivation to investigate how trust is created, specifically how altruistic campaigners for a better world can harness trust for their efforts.

Possibilities include building networks through personal contact, personally connecting with a range of people from diverse sectors in society, developing trusted communication channels, learning how to avoid rushing to judgement, and identifying ways to verify information using multiple channels. It should be possible to probe prior campaigns to explore how trust is built and undermined, and apply insights to a world with deepfakes.

Given the rapid development of deepfake technology, lessons will need to be continually updated. This puts a premium on continual learning.

Preparing for AI

Killer robots and deepfakes are just two of the many impending impacts of AI. Others include facial recognition and social media analysis. Some of these, like killer robots, are unlikely to affect activists, whereas others, like deepfakes, potentially will require major adjustments. Is there some way to anticipate and get ahead of the impacts of new technology?

Technologies are not autonomous (Winner 1977): they are developed by humans, most commonly for specific purposes. Most are developed with good intentions, but people often disagree about these intentions, for example whether 'better' weapons are a good idea. In principle, activists should be engaging in the innovation process, putting forward their values in helping decide research and development priorities (Sclove 1995). However, despite the efforts of campaigners to democratise the processes of technological innovation, in practice the main players are powerful governments and corporations, with most of the population positioned as users, as consumers. Most activists have little input into development and promotion, so their main choices are about which technologies to use personally and which ones to campaign for or against. In many cases, it is mainly a question of adapting and responding to new technologies, and so far that is the main way of interacting with AI.

There is a role for activists, individually and in groups, to learn about AI and its applications, to campaign when appropriate, to foster greater understanding, and to prepare and adapt when necessary. Going beyond this, there is a greater challenge: becoming involved in helping set agendas for innovation. That is already the case with environmental and energy technologies, where campaigners have played a major role, for example in championing renewable energy. So why not do the same with AI?

Acknowledgements

Thanks to Jungmin Choi, Kelly Gates, Tim Johnson-Newell, Olga Kuchinskaya and Julia LeMonde for helpful feedback.

References

Baker, D. 2022 *Should We Ban Killer Robots?* Polity Press: Cambridge.

Fallis, D. 2021 'The epistemic threat of deepfakes', *Philosophy and Technology,* 34, 4: 623–643.

Frankel, P. 2001 *An Ordinary Atrocity: Sharpeville and its Massacre*, Yale University Press: New Haven, CT.

Grothaus, M. 2021 *Trust No One: Inside the World of Deepfakes*, Hodder Studio: London.

Karnouskos, S. 2020 'Artificial Intelligence in digital media: the era of deepfakes', *IEEE Transactions on Technology and Society*, 1, 3: 138–147.

Monkeypaw Productions and BuzzFeed 2018 'You won't believe what Obama says in this video!' YouTube, https://www.youtube.com/watch?v=A7BZKj3Cyh4 (accessed 14/12/2022).

Rosendorf, O., Smetana, M. and Vranka, M. 2022 'Autonomous weapons and ethical judgments: experimental evidence on attitudes toward the military use of "killer robots"', *Peace and Conflict: Journal of Peace Psychology*, 28, 2: 177–183.

Scharre, P. 2018 *Army of None: Autonomous Weapons and the Future of Wa*r, Norton: New York.

Sclove, R. E. 1995 *Democracy and Technology*, Guilford Press: New York.

Skitz4twenty 2016 'The Rock singing to Hillary', YouTube, https://www.youtube.com/watch?v=A7BZKj3Cyh4 (accessed 14/12/2022).

Winner, L. 1977 *Autonomous Technology: Technics-out-of-control as a Theme in Political Thought*, MIT Press: Cambridge, MA.

Author
Brian Martin is emeritus professor of social sciences at the University of Wollongong. He is the author of 22 books and hundreds of articles on nonviolent action, dissent, scientific controversies, tactics against injustice, and other topics.