Winning the Battle, Losing the War

Ryan Jenkins shows how blowback against apparently successful technologies can render them counterproductive

At least since the Industrial Revolution, humanity has had a troubled relationship with technology. Even while standards of living have skyrocketed and life expectancies lengthened, we have often been shocked or dismayed by the unforeseen disruptions that technology brings with it. I suspect that a major source of this myopia is the naivete of one popular view of technology: that technologies are merely neutral tools and that our engagements with particular technologies are episodic or, in the words of Langdon Winner, *brief, voluntary, and unproblematic*. I think view is simplistic – and appreciating both the consequences of longer-term technological *policies* and the interplay between a technology and its *social context* can help us anticipate these negative consequences. In particular, we should appreciate how even an efficient and reliable technology can nurture social circumstances that will undermine the very goals that technology is meant to serve.

Of course, the introduction of new technologies often helps us accomplish our ends. For example, the car dramatically increased our mobility, allowing us to travel farther quicker. But at the same time, it catalysed other changes that ultimately thinned out our communities by incentivising the creation of sprawl and suburbs. While the car enabled us to get from A to B more quickly, at the same time, it made it easier for urban planners to build A and B father apart from one another.

A more nuanced philosophy of technology can help us appreciate this problem: that technology can accomplish some goal we have when that goal is understood narrowly – the car improved mobility. But because of its effects on its social and cultural environment, the very same technology can actually undermine our larger projects – the car undermined closeness and community. In short, technologies often help us win a battle but lose the larger war of which those battles are a part.

Consider another example. America's policy of using drones for targeted assassination has attracted vociferous criticism. But some scholars have put forth compelling arguments that there is nothing especially problematic about the use of drones as weapons of war – nothing that makes them importantly different from, say, long-range artillery. And others have assembled compelling data that show that drones actually cause fewer collateral deaths, per strike, than other methods of war, such as using special forces.

If it's true that drones kill fewer civilians with each individual strike, shouldn't that be the end of the moral discussion? Maybe not – let's consider the effects of the technology on the longer-term project of winning a war. Even if drones kill fewer civilians in each individual strike, *a policy of drone strikes* could be morally flawed because it leads to a "low boil" of warfare that can lead to more civilian casualties over time. This is because drone strikes, as a weapon of war, inspire a uniquely intense resentment, such that *civilian deaths by drone strike are more likely than deaths by other means to inspire civilians to become insurgents*. This phenomenon is called *blowback*.

Thus, even though each drone strike may cause fewer civilian deaths than other methods of war, if civilian deaths by drone strike are more likely to inspire others to become insurgents, then this can perpetuate counter-terror operations, demanding more drone strikes over the long run. While each one of these strikes may kill fewer civilians than other methods of war, they also threaten a vicious, self-feeding cycling, leading to greater *total* civilian deaths over the lifetime of the counter-insurgency campaign.

So, the technology may allow us to win more battles, but in the same stroke make it harder for us to win the war. Drones may succeed at a particular goal when that goal is

described narrowly: kill the guilty and spare the innocent. But they actively undermine a broader goal, which is to end the war while killing as few of innocent people as possible overall. At least, this is the worry.

A lot of our discussions about technology in the Western world are bedevilled by the view called *technological instrumentalism*: that technologies themselves are morally neutral instruments that merely help us achieve some goal more efficiently than before. They can be put to beneficial or nefarious ends, of course, but nonetheless they themselves remain inert. When we examine technologies in this way, we miss the broader, more subtle ways that technologies, once deployed into a social context, can exert powerful forces on our behaviours and beliefs.

We should appreciate how even an efficient technology can nurture social circumstances that undermine the very goals that technology is meant to serve

Decades ago, Emmanuel Mesthene offered a treatment of how technologies alter the network of incentives in society and, in so doing, can dramatically reshape human life. His explanation went like this: Social institutions are constructed to achieve goals. But social institutions do not adopt new technologies in a vacuum. Instead, old ways are often replaced, or institutions are reconfigured to make more effective use of the new technologies. In the process of this reconfiguration, other valuable purposes that the institution served can be undermined. As a result, technologies often incentivise some valuable goals while, at the same time, causing other goals to be neglected. Thus, Mesthene's famous dictum, that technology "*has both positive and negative effects, and it usually has the two at the same time and in virtue of each other*" (his emphasis). We don't need to adopt some mystical position that says that technologies are intrinsically moral or immoral, but instead we can say that the disruptive effects of a technology arise out of the technology's interplay with pre-existing social practices.

Take another example that's closer to home. Consider not military technologies, but the artificially intelligent (AI) prediction technologies that are currently used by more than 60 of America's police departments. This is called "predictive policing": the practice of using AI to predict crimes. Often, this involves predicting their timing and location, but it can also mean identifying individuals who are thought to be at high risk of committing crimes.

The most common objection to predictive policing is that it is a technologically-disguised version of racial profiling. Researchers have accumulated extensive examples of purportedly objective algorithmic systems generating biased verdicts. That's a problem that's worth investigating, for sure.

But I am interested in a different kind of objection to predictive policing, one that will stand even if there is nothing intrinsi-

cally wrong with using computer models to patrol neighbourhoods or identify people who might commit crimes, and even if predictive policing methods are no more biased than other methods of policing.

Once again, if we're supposing there's nothing intrinsically wrong with the technology, isn't it game over for critics of predictive policing? Well, not necessarily, because there might be negative consequences that emerge only when a long-term policy of predictive policing is in place. In particular, such a policy might erode trust in a way that undermines public safety. Notice the parallels to the use of drones: what is successful in a single instance may be counterproductive as a policy.

All of this depends on how we understand and conceptualise the goal of technologies in the first place. We might think, at first glance, that the goal of these technologies is to accurately predict when and where crimes will take place. And there is reason to believe they are successful at doing that. But this goal is only a part of the larger goal of the criminal justice system, which is securing convictions that remove criminals from the streets and deter others from committing crimes. In short, this technology might accomplish its goal when that goal is narrowly conceived, but undermine the larger goal it is ultimately meant to serve.

Even though there is evidence from several communities that predictive policing technologies do reduce crime, we have to turn to the nature of the technology to see its potential to become a double-edged sword. The key point is that the use of these technologies nurtures suspicion and hostility between police and the communities they serve. While there is no direct empirical evidence of this claim, there is good evidence for closely related claims concerning police-community relations and trust of algorithms in general.

First, we know that the populace at large, and in particular minority communities, harbour an acute suspicion of the use of algorithms for criminal justice purposes. Pew found in 2018 that 58% of Americans worried that machine learning algorithms will always reflect some amount of human bias. Pew found widespread anxiety about the use of machine learning algorithms in making predictions about the likelihood of a criminal recidivating – so-called "criminal risk assessments" or "risk scores". These predictions are often used in judicial or parole hearings to determine the length and nature of a defendant's sentence. Pew found that 56% of Americans found this use of algorithms "unacceptable". And the percentage of people who found risk scores to be unfair rose to 61% for black respondents.

Even if drones kill fewer civilians in each individual strike, a policy of drone strikes could be morally flawed

Second, to the extent that law enforcement relies on the cooperation and support of local communities, we should expect the criminal justice system to meet with greater difficulties where the operations of police are informed by artificially intelligent predictions. Police departments who adopt


© Victor Garcia

predictive policing technologies do so to increase their efficiency at apprehending and deterring criminals through increased police presence, i.e. by being in the right place at the right time. However, police enforcement efforts are just one element of the broader criminal justice system, which is best conceived as a network of interrelated activities. Before crimes are committed, it includes intelligence gathering, receiving tips, and observation. After crimes are committed it includes community-generated crime reports, evidence collection, investigation, deposing witnesses, securing testimony, and convincing juries to convict.

Many of these functions rely on the trust of the communities being policed. Declining community trust makes police officers hesitant to conduct stops when they believe it is necessary, which in turn increases the threat to public safety. Declining trust can also reduce "collective efficacy", the ability of communities to contribute to their own safety. For example, citizens who distrust the police or are sceptical of their methods might be less likely to cooperate with police, including reporting crimes, cooperating with investigations, offering depositions, or testifying at trial. Juries who are sceptical or distrustful of the use of algorithms might be less likely to convict a defendant if a crucial aspect of the state's case – for example, if the "probable cause" for a search – relied on the verdicts of an algorithm. Thus, while the police's ability to make predictions about the timing and location of crimes could be augmented by the use of machine learning algorithms, *the very same technology* could undermine many of the *other elements* of the criminal justice process that are necessary to secure convictions.

With fewer convictions secured, the criminal justice system's crucial goal of *imprisoning criminals* would be undermined. And since the deterrent effect of law enforcement relies on the ability of the state to secure convictions – rather than merely apprehend criminals – the criminal justice system's goal of deterring potential criminals would also be undermined.

AI prediction technologies are used by more than 60 of America's police departments – it's called "predictive policing"

This is technological blowback. To the extent that the successful execution of justice relies on the trust and understanding of policed communities and the broader public, and to the extent that the widespread use of predictive algorithms in pursuit of justice imperils this trust and understanding, the use of predictive policing could ultimately result in an overall reduction in the ability of the state to fight crime. Winning the battles, losing the war.

Predictive policing threatens to have this corrosive effect on law enforcement because interventions based on algorithmic predictions undermine public trust more significantly than interventions based on human judgement. The results above from Pew suggest that the use of algorithms in

criminal justice is *an independent source of distrust* among the public – and among minorities in particular. If these speculations are correct, then Mesthene's framework is vindicated once again: supposing that algorithmic predictions used in policing are accurate, the adoption of this technology has clear benefits in terms of public safety. However, the very same technology – its same technological features – could become a source of pronounced resentment among the communities that are being policed by artificial intelligence. Thus, predictive algorithms could plausibly have good and bad effects, *at the same time, and in virtue of each other*.

You might think that this distrust in algorithms is misplaced as long as the predictions are accurate. But it seems to me that this distrust, even if mistaken, is problematic because it has real consequences for the effectiveness of law enforcement in fighting crime. (Never mind that minority populations often *do* have good reasons to be suspicious of law enforcement.) Because this distrust is more acute among minority populations, they would likely be less willing to assist police in testifying or informing on criminals. The result is that *those who victimise minority neighbourhoods benefit especially* from this reduced trust, which produces a disproportionate negative impact on those neighbourhoods.

This blowback to predictive policing technologies should be considered inseparable from the technology itself. Or, rather, *only conceptually separable* – but in such a way that the separation is irrelevant for policy-making. When the effects of a technology and arise inextricably from the combination of its design and its social context, it is worth considering the more provocative view that *the technology itself has moral properties*.

To consider the technology to be successful as long as it's making correct predictions would be a mistake. It would be to artificially sever the connection between a technology and its environment, to amputate it from the living social system in which it operates. Decisions to deploy technologies rarely – if ever – concern individual instances. Economic, bureaucratic, and practical pressures force institutions to adopt policies rather than individual actions. We should be at least as concerned about the consequences of *technological policies* as we are about the individual uses of a technology. If there is any hope for imposing order on the seeming chaos of technological disruption, for anticipating the full range of consequences that technologies cause, we must investigate the entire network of social effects they bring along with them – including ultimately the paradoxical possibility that they could undermine their own success.

Ryan Jenkins is an associate professor of philosophy at California Polytechnic State University in San Luis Obispo and a co-editor of recent books Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence *(Oxford 2017) and* Who Should Die?: The Ethics of Killing in War *(Oxford 2017). He investigates how technologies enable or encumber meaningful human lives.*

This research was funded by NSF award #1917707 shared between Cal Poly, San Luis Obispo and the University of Florida.