PATRYK HARDZING

# AI, *warbot*

Artificial intelligence is set to rewrite the rules of warfare in subtle and terrifying ways, says Kenneth Payne

"ONLY the dead have seen the end of war," the philosopher George Santayana once bleakly observed. Our martial instincts are deep-rooted. Our near relatives chimpanzees fight "total war" that sometimes leads to the annihilation of rival groups of males. Archaeological and ethnographical evidence suggests that warfare among our hunter-gatherer ancestors was chronic.

Over the millennia, we have fought these wars according to the same strategic principles based in our understanding of each other's minds. But now we've introduced another sort of military mind – one that even though we program how it thinks, may not end up thinking as we do. We're only just beginning to work through the potential impact of artificial intelligence on human warfare., but all the indications are that they will be profound and troubling – in ways that are both unavoidable and unforeseeable.

We're not talking here about the dystopian sci-fi trope of malign, humanoid robots with a free rein and a killer instinct, but the far more limited sort of artificial intelligence that already exists. This AI is less a weapon per se, more a decision-making technology. That makes it useful for peaceful pursuits and warfare alike, and thus hard to regulate or ban.

This "connectionist" AI is loosely based on the neural networks of the human brain. Networks of artificial neurons are trained to spot patterns in vast amounts of data, gleaning information they can use to optimise a "reward function" representing a specific goal, be that optimising clicks on a Facebook feed, playing a winning game of poker or Go, or indeed winning out on the battlefield.

In the military arena, swarms of autonomous drones are already deployed from pods on aircraft, and autonomous software can manoeuvre vehicles with increasing dexterity. In the air – in simulators at least – it has outfought skilled pilots. There are systems that scan hours of imagery looking for targets, that automatically respond to incoming missile threats, that prioritise information for human pilots and that shift radar bands in a lightning-fast battle of detection and deception.

This raises obvious, much discussed ethical questions. Can AI systems really know who to target? Shouldn't humans have the final say in life-or-death decisions? But the implications for how war is prosecuted – for strategy – have been less widely explored. To understand how profound they are, we must first understand strategy's very human underpinnings.

Social intelligence gives humans a powerful advantage in conflict. In war, size matters. Victory generally goes to the big battalions, a logic described in a formula derived by the British engineer Frederick Lanchester from studies of aerial combat in the first world war. He found that wherever a battle devolves to a melee of all against all, with ranged weapons as well as close combat, a group's fighting power increases as the square of its size.

That creates a huge incentive to form ever-larger groups in violent times. Humans are good at this, because we're good at understanding others. We forge social bonds with other unrelated humans, including with strangers based on ideas, not kinship. Trust is aided by shared language and culture. We have an acute radar for deception, and a willingness to punish non-cooperating free-riders. All these traits have allowed us to assemble, organise and equip large and increasingly potent forces to successfully wage war.

Social intelligence also allows weaker, smaller groups to stave off defeat. The use of deception, fortification and terrain and disciplined formations all can offset the advantages of scale and shock. In the film *300*, crack Spartan troops at one point charge headlong into the vastly outnumbering Persian army at the Battle of Thermopylae. In reality, that would spell disaster. As the Ancient Greek historian Herodotus relates, the Spartans stuck to using the narrow confines of the mountain pass, arranged into a disciplined formation with interlocked shields to hold off the Persians. This too is strategic intelligence.

Underlying it is theory of mind – the human ability to gauge what other humans are thinking and how they will react to a given situation, friend or foe. The ancient Chinese strategist Sun Tzu counselled leaders to know themselves and know their enemies, so that in 100 battles they would never be defeated. Theory of mind is essential to answer strategy's big questions. How much force is enough? What does the enemy want, and how hard will they fight for it?

Strategic decision-making is often instinctive and unconscious, but also shaped by deliberate reflection and an attempt at empathy. This has survived even into the

›

nuclear era. Some strategic thinkers held that nuclear weapons changed everything because their destructive power threatened punishment against any attack. Rather than denying aggressors their goals, they deterred them from ever attacking.

That certainly did require new thinking, such as the need to hide nuclear weapons, for example on submarines, to ensure that no "first strike" could destroy all possibility for retaliation. Possessing nuclear weapons certainly strengthens the position of militarily weaker states; hence the desire of countries from Iran to North Korea to acquire them.

But even in the nuclear era, strategy remains human. It involves chance and can be emotional. There is scope for misperception and miscommunication. And a grasp of human psychology can be vital for success.

## What are you thinking?

Take the Cuban missile crisis, an event intensely studied by psychologists and strategists since. In 1962, US President John F. Kennedy was given alarming evidence that the Soviet Union was positioning nuclear missiles on Cuba. His immediate reaction was anger, and a desire to strike out militarily, even at the risk of escalating the cold-war conflict. But that soon gave way to a deliberate, reflective attempt to empathise with Nikita Khrushchev's blustering. The Soviet leader had tried to bully Kennedy at their first meeting, and during the crisis sent first an emollient letter, then a tougher one. Kennedy crafted a solution that, crucially, saved Khrushchev's face: in a tense stand-off, social intelligence and theory of mind were decisive.

Artificial intelligence changes all this. First, it swings the logic of strategy decisively towards attack. AI's pattern recognition makes it easier to spot defensive vulnerabilities, and allows more precise targeting. Its distributed swarms are hard to kill, but can concentrate rapidly on critical weaknesses before dispersing again. And it allows fewer soldiers to be risked than in warfare today.

This all creates a powerful spur for moving first in any crisis. Combined with more accurate nuclear weapons in development, this undermines the basis of cold-war nuclear deterrence, because a well-planned, well-coordinated first strike could defeat all a defender's retaliatory forces. Superior AI capabilities would increase the temptation to strike quickly and decisively at North Korea's small nuclear arsenal, for example.

"AI does not experience the emotion and empathy felt by human strategists"

By making many forces such as manned aircraft and tanks practically redundant, AI also increases uncertainty about the balance of power between states. States dare not risk having second-rate military AI, because a marginal advantage in AI decision-making accuracy and speed could be decisive in any conflict. AI espionage is already under way, and the scope for a new arms race is clear. It's difficult to tell who is winning, so safer to go all out for the best AI weapons.

Were that all, it would be tempting to say AI represents just another shift in strategic balance, as nuclear weapons did in their time. But the most unsettling, unexplored change is that AI will make decisions about the application of force very differently to humans. AI doesn't naturally experience emotion, or empathy, of the sort that guides human strategists such as Kennedy.

We might attempt to encode rules of engagement into an AI ahead of any conflict – a reward function that tells it what outcome it should strive towards and how. At the tactical level, say with air-to-air combat between two swarms of rival autonomous aircraft, matching our goals to the reward function that we set our AI might be doable. Win the combat, survive, minimise civilian casualties – such goals translate into code, even if there may be tensions between them.

But as single actions knit together into military campaigns, things become much more complex. Human preferences are fuzzy, sometimes contradictory, and apt to change in heat of battle. If we don't know exactly what we want, and how badly, ahead of time, machine fleets have little chance of delivering those goals. There is plenty of scope for our wishes and an AI's reward function to part company. Recalibrating the reward function takes time, and you can't just switch AI off mid-battle – hesitate for a moment, and you might lose.

That's before we try to understand how the adversary may respond. Strategy is a two-player game, at least. If AI is to be competitive, it must anticipate what the enemy will do.

The most straightforward approach, which plays to AI's tremendous abilities in pattern recognition and recall, is to study an adversary's previous behaviour and look for regularities that might be probabilistically modelled. This method was used for example by AlphaGo, the DeepMind AI that beat the human champion Lee Sedol at the board game Go in 2016. The Go board represents a large, yet still limited, "toy universe" with a vast array of possible future moves. Yet given its opponent's likely response, the machine can narrow the search to the moves most likely to lead to victory, and then work out a winning course of action – all at blinding speed.

With enough past behaviour to go on, this works even in a game such as poker where, unlike Go, not all information is freely available and a healthy dose of chance is involved: AI can now beat world-class poker players when it plays them repeatedly.

PICTURE CREDIT

Autonomous drones are already deployed militarily, here on the Iraq-Turkey border

John F. Kennedy's human reactions were decisive in solving the 1962 Cuban missile crisis

PICTURE CREDIT

This approach could again work well at the tactical level – anticipating how an enemy pilot might respond to a manoeuvre, for example. But it falls down as we introduce high-level strategic decisions: there is too much unique about any military crisis for previous data to model it.

An alternative approach is for an AI to attempt to model the internal deliberations of an adversary. But this only works where the thing being modelled is less sophisticated, as when an iPhone runs functional replicas of classic 1980s arcade games. Our strategic AI might be able to intuit the goals of an equally sophisticated AI, but not how the AI will seek to achieve them. The interior machinations of an AI that learns as it goes are something of a black box, even to those who have designed it.

Where the enemy is human, the problem becomes more complex still. AI could perhaps incorporate commonplace themes of human thinking, such as the way we systematically inflate low-risk outcomes. But that's AI looking for patterns again. It doesn't understand what things mean to us; it lacks the evolutionary logic that drives our social intelligence. When it comes to understanding what others intend – "I know that you know that she knows" – machines are not at the races.

Does that matter? Humans aren't infallible mind-readers, and in the history of international crises misperception abounds. In his sobering account of nuclear strategy, *The Doomsday Machine*, Daniel Ellsberg describes the original US early warning system signalling an incoming Soviet strike. In fact, the system's powerful radar beams were echoing back from the surface of the moon. Would a machine have paused for thought to ascertain that error before launching a counterstrike, as the humans involved did?

"We can't bury our heads and say it won't happen – the technology already exists"

Humans try to reason about what adversaries want, and understand that within the context of their own experience, motivations and emotions. Machines might not share Kennedy's emotional knee-jerk response in 1962, but they also don't share his capacity to reflect on his adversary's perspective.

And their own moves are often unexpected. In its second game against Lee, AlphaGo made a radical move wholly unexpected by onlooking human experts. This wasn't remarkable creativity or a searing insight into Lee's game plan. The game-winning "move 37" was down to probabilistic reasoning and a flawless memory of how hundreds of thousands of earlier games had played out.

The last thing humanity needs is a blindingly fast, offensively brilliant AI that makes startling and unanticipated moves in confrontation with other machines.

And there won't necessarily be time for human judgement to intercede in a battle of automatons before things gets out of hand. At the tactical level, keeping a human in the loop would ensure defeat by faster all-machine combatants. Despite the stated intentions of liberal Western governments, there will be ever-less scope for human oversight of blurringly fast tactical warfighting.

## Cold probabilities

The same may be true at more elevated strategic levels. Herman Kahn, a nuclear strategist on whom the character Dr Strangelove was partly based, conceived of carefully calibrated "ladders" of escalation. A conflict is won by dominating an adversary on one rung, and making it clear that you can suddenly escalate several more rungs of intensity, with incalculable risk to the enemy – what Kahn called "escalation dominance".

In the real world, the rungs of the 'ladder' are rather imprecise. Imagine two competing AI systems, made of drones, sensors and hypersonic missiles, locked in an escalatory game of chicken. If your machine backs off first, or even pauses to defer to your decision, it loses. The intensity and speed of action pushes automation ever higher. But how does the machine decide what it will take to achieve escalation dominance over its rival? There is no enemy mind about which to theorise; no scope for compassion or empathy; no human to intimidate and coerce. Just cold, inhuman probabilities, decided in an instant.

That was move 37 of AlphaGo's second game against the world champion. Perhaps it is also early December 2041, and a vast swarm of drones skimming over the ocean at blistering speed, approaching the headquarters of the US Pacific Fleet. We can't bury our heads and say it won't happen, because the technology already exists to make it happen. We won't be able to agree a blanket ban, because the strategic advantage to anyone who develops it on the sly would be too great. The solution to stop it happening is dispiritingly familiar to scholars of strategic studies – to make sure you win the coming AI arms race. ■

Kenneth Payne is at the School of Security Studies, King's College London. He is author of *Strategy, Evolution and War: From apes to artificial intelligence* (Georgetown University Press, 2018)