Routledge
Taylor & Francis Group

Survival

Global Politics and Strategy

ISSN: 0039-6338 (Print) 1468-2699 (Online) Journal homepage: https://www.tandfonline.com/loi/tsur20

Artificial Intelligence and Nuclear Command and Control

Mark Fitzpatrick

To cite this article: Mark Fitzpatrick (2019) Artificial Intelligence and Nuclear Command and Control, Survival, 61:3, 81-92, DOI: 10.1080/00396338.2019.1614782

To link to this article: https://doi.org/10.1080/00396338.2019.1614782

Artificial Intelligence and Nuclear Command and Control

Mark Fitzpatrick

In early June, a series of alarming reports roiled relations among the major powers, sparking concerns of a nuclear war. A low-level conflict between Russian and US special-operations forces in Syria apparently left three American soldiers dead from nerve gas. YouTube videos of family members of cabinet officials and senators hurriedly fleeing Washington, and social-media accounts of missile silos in the prairie states going on high alert, triggered Moscow's strategic-warning systems to warn the Kremlin leadership that a US missile launch could be imminent. Separately, satellite imagery appeared online of the South China Sea that utilised new, cutting-edge technology rendering the oceans effectively 'transparent' and revealed the location of three Chinese submarines. Assuming a US government connection, Chinese leaders saw it as a sharp escalation in Washington's presumed encirclement strategy. Meanwhile, management software in several Chinese nuclear power plants generated false-positive safety reports that triggered shutdown decisions. Some wondered if this, too, was due to adversarial misfeasance. Luckily, officials in all countries soon came to realise that the culprit in most of these actions was a shadowy non-governmental organisation seeking to spark conflict through falsified information on social media and the exploitation of private-sector technology breakthroughs.

Mark Fitzpatrick is an IISS associate fellow, was executive director of IISS–Americas from 2015 through 2018, and headed the IISS Non-Proliferation and Nuclear Policy Programme for 13 years. He had a 26-year career in the US Department of State, including as deputy assistant secretary of state for non-proliferation.

Survival | vol. 61 no. 3 | June–July 2019 | pp. 81–92 DOI 10.1080/00396338.2019.1614782

The year was 2021. The fictionalised events occurred in an IISS tabletop exercise in London in November 2018, for a project funded by the Carnegie Corporation of New York designed to examine the potential impact of artificial intelligence (AI) on nuclear strategic stability. Experts from China, Russia, the United States, the United Kingdom and the Czech Republic played roles as decision-makers in Beijing, Moscow and Washington as they sought to understand, address and take advantage of the scenarios and variations posed by the IISS control team.

The major powers' nuclear command-and-control systems increasingly rely on AI programmes, or, more precisely, expert systems and machine-learning algorithms, to enhance information flow, situational awareness and cyber security. Such capabilities can provide such systems with a larger window of opportunity in which to respond in the event of a crisis and thereby support de-escalation. Malevolent actors, however, can also use new technologies offensively to deceive, disrupt or impair command-and-control systems and their human controllers. The IISS tabletop exercise sought to explore several of these malfeasance pathways and vulnerabilities in plausible crisis scenarios. It demonstrated how an AI arms race could reduce strategic stability as the nuclear-weapons states become more reliant on AI for strategic warning in relation to nuclear command-and-control functions.[1]

Escalation via 'deep fakes'

A key danger played out in the exercise was the potential for third parties to spoof warning systems and embed disinformation to fool human operators. In the scenario, a hitherto unknown non-state actor, the World Peace Guardians, circulated falsified videos and photographs on social-media platforms to create the impression that three US special-operations-forces soldiers had been killed by nerve gas in clashes with Russian military advisers in Syria. Some US pundits argued the legal case for the use of tactical nuclear weapons in response. Doctored videos then appeared on American and Chinese media platforms that showed the families of several prominent US officials hurriedly departing Washington DC. Further eyewitness accounts on social media claimed that missile silos in the western US had gone to high alert, and that the crews of two had even opened the silo

doors. As a result of these and other secondary indicators used by Russian AI-driven situational-awareness algorithms, Moscow's strategic-warning systems began informing the Kremlin leadership that a US missile launch could be imminent. In the first situation report generated by the control team, Chinese President Xi Jinping cautioned the US not to conduct a nuclear strike against any country and warned that if the US did not provide 'credible evidence' that it was not mobilising for war, China would have to take unspecified defensive actions.

Generally speaking, the scenario was a plausible example of escalation by third parties. In this case, a non-state actor made a deep fake sufficiently believable that it generated a crisis between two nuclear states. A workshop background paper explained how offensive AI capabilities could widen the psychological distance between the attacker and its target. The paper forewarned that a third-party actor might attempt to use AI-driven adversarial inputs, data-poisoning attacks, and audio and video manipulation to create escalatory effects between nations. While other early-warning systems would eventually discredit the spoof, it would still create high levels of uncertainty and tension in a short period of time. Doctored videos would likely force both parties to put their respective militaries on high alert. They would utilise overhead imagery, signals intelligence or human reporting to determine the reality on the ground. Collecting and processing the intelligence would take precious time during an escalatory crisis in which AI algorithms were urging immediate response actions.

In round two of the exercise, set two days later, Russian and US conventional military forces were on alert and beginning to prepare for possible contingencies. Russian Tu-160 Tupolev supersonic, nuclear-capable bombers were airborne and conducting flights over international waters in the Bering Strait and northern Atlantic Ocean, and US satellite imagery identified numerous Russian transport-erector-launcher units moving within Russian territory. Several US B-52H *Stratofortress* nuclear-capable bombers were airborne and US missile silos were on alert. US AI-driven situational-awareness algorithms downplayed the gravity of the situation, however. Another AI system assessed that the viral images of asphyxiated US soldiers were inauthentic, based on their pixilation and online propagation.

Meanwhile, though, many US citizens were spontaneously departing major cities on the eastern seaboard, overwhelming transportation routes.

In the denouement of round three, trilateral information sharing conducted by the respective computer emergency-response teams and law-enforcement agencies of China, Russia and the US concluded that the images, videos and other social-media postings by the World Peace Guardians organisation were falsified. Both Russia and the US publicly acknowledged that no chemical-weapons attack had occurred in Syria. As a result, both countries' militaries returned to standard operating procedures. Public fears of nuclear war and mass evacuations of cities also subsided. Yet relations between Russia and the US, in particular, remained testy at best.

Skewed early-warning assessments

Deep fakes by third parties can be magnified by AI capabilities that fall into the wrong hands and are used to generate false positives. In our scenario, three months before the crisis and start of play, US Cyber Command's 'Unified Platform' for managing and coordinating integrated cyber-, electronic- and information-warfare operations was providing statistically anomalous outputs regarding situational-awareness assessments related to the early warning of selected advanced persistent threat datasets linked to Russian cyber actors. The questionable reports appeared to be skewed by mathematical coefficients derived from large-scale metadata analysis during the beta-testing training phase of the Unified Platform's software.

A workshop background paper noted that AI programmes are driven by the data that they receive – the digital equivalent of the adages 'you are what you eat' and 'garbage in, garbage out'. This is also true of the process by which such programmes initially formulate their pattern-recognition models and evaluative procedures for use in future contexts. The malign manipulation of input data can not only pervert the output of AI functions in specific instances, but also undermine the reliability of an entire algorithm if accomplished during the 'training' phase for such programmes. Our scenario showed vividly how this could play out. Non-malign, human-design factors can also corrupt AI assessments. Because human beings define an AI system's algorithms and curate its training data, they can unintentionally

insert their own biases into the system. This can cause the system to behave in unintended ways that may be undetectable to its operators due to the feedback loop created by those very biases. In addition, nuclear command-and-control personnel also faced a black-box problem in determining how or why a system came to a certain conclusion. While system inputs and outputs can be observed, the speed and scale of system processes make it difficult for personnel to isolate the logic behind any particular prediction. Once the operation of AI systems is triggered, humans are unable to monitor the systems' decision calculus in real time.

In the exercise, the US team was aware of this range of potential problems with respect to AI-driven assessments provided by Cyber Command's Unified Platform. They worried that Russian strategic-warning systems might be similarly skewed. In particular, the Americans realised that the false positives produced by Russia's strategic-warning system need not have been caused by a malign actor, as they could also have resulted from a problem intrinsic to the algorithm. To preclude Russia from perceiving any US effort to investigate as an attempt to manipulate Russia's early-warning and command-and-control systems, the US team decided not to conduct any reconnaissance. In turn, the scenario in round two had Kremlin officials learning of a prior intelligence operation by the Main Directorate of the General Staff of the Russian Armed Forces (GRU), Russia's military-intelligence agency, injecting skewed training data into AI-driven situational-awareness algorithms of the US military during their developmental phase. The objective was to prevent American AI systems from being able to recognise Russia as an aggressor.

The background papers for the exercise had noted how data-poisoning attacks introducing skewed inputs into the training dataset for an AI machine-learning process can distort an AI system's output by degrading its ability to distinguish between good data and bad data. These types of operations are usually carried out through techniques known as content generation, feedback weaponisation, perturbation injection and man-in-the-middle attacks. However, an attacker usually will not have direct access to the actual training data. To overcome this obstacle, the attacker will target the actors or methods used to collect and store machine-learning training

data, such as cloud graphics-processing units, web-based repositories, and contractors or third-party service providers.

Understanding that early-warning assessments from its own AI-driven software could be skewed, the Russian team chose largely to ignore the AI algorithms that were producing results so disparate from the reality they otherwise perceived, while also deciding not to reveal any analytic problems they were encountering to their adversaries. Unaware that the Russians had, in effect, pulled the plug on the AI-driven early-warning system, the US team spent considerable time trying to persuade their counterparts to investigate its flaws. The US team also considered offering to work with Russia to neutralise any malicious third-party-introduced software problems. They recognised, however, that any such effort might itself be seen as an attempt to compromise the Russian command-and-control system. Indeed, the Russians reacted defensively. With team members playing to stereotype, trilateral meetings and a session of the United Nations Security Council ended inconclusively amid mutual recriminations.

False-positive safety alerts

In the notional ten months leading up to our crisis scenario, four civilian nuclear power plants operating in densely populated Chinese provinces began registering false-positive safety alerts by their AI-driven management software. In each case, sensors reported structural concerns in the facility that led supervising authorities to enact emergency-management procedures that stopped operations, although engineering tests revealed that no structural damage or harm to humans had actually occurred. Even so, the repeated stoppages in electricity production by these important power plants were having adverse economic effects. The operating authorities accordingly requested that the AI software provider, named GaiaForce, adjust the sensing functions to require a much higher threshold of structural-integrity damage before triggering automatic shutdown protocols. This allowed all of the power plants in question to operate at or near full capacity. GaiaForce was unable to find any defects in the functionality of its code.

Left unclear during the initial rounds was whether the false positives resulted from malicious activities. The workshop background paper

explained how 'raising the noise floor' – launching a cyber attack perpetrated via digital noise and extraneous inputs containing minor elements of an actual threat – can cause an AI system to generate a stream of false positives. These false positives can lead operators to reconfigure the AI's machine-learning algorithm to avoid this error in the future. At that point, the adversary can launch an attack through the same method that the system was reconfigured to ignore. This type of attack involves social engineering in that, unlike other technical-intelligence operations, it targets the human operators and induces them to effect a change that favours the offensive adversary. While the AI system faithfully does what it has been reinstructed to do, the parameters no longer successfully serve their original defensive purposes. Thus, in round two, a nuclear power plant in Hubei province suffered structural damage as a result of operating beyond engineering parameters due to the loosening of the security indicators contained in the GaiaForce management software. Although the situation was contained in time, safety experts claimed that they were 'minutes away from the next Chernobyl or Fukushima'.

In round three, China provided detailed technical data of operational anomalies to the International Atomic Energy Agency (IAEA), along with portions of the GaiaForce source code. After obtaining this information, the US National Security Agency (NSA) assessed that the GaiaForce source code identified the potential for additional 'remote administration features' that could arise if that software were run on systems that were also infected by a malware exploit that the NSA determined was devised by the GRU. The NSA further judged that the GRU had created the GaiaForce source code and used it in its most sensitive cyber operations. According to the NSA director, the combination of GaiaForce and the malware could be used to 'weaponise' civilian nuclear reactors. That this did not occur in the exercise was down to a combination of luck and Chinese willingness to cooperate with other concerned actors.

Hijacking private-sector technology

Among the most revealing vulnerabilities exposed in the exercise was the way in which private-sector technological breakthroughs might be expropriated by another non-state actor for malevolent purposes, with dire

consequences for strategic stability. In the hypothetical two months before the exercise started, a US technology company called QuantumAI, which received significant seed funding from the CIA and won Defense Advanced Research Projects Agency contracts for its advanced magnetometers and gravimeters, partnered with another company to launch four open-source, quantum-sensing satellites that used AI to measure infinitesimal changes in magnetic and gravitational fields on the earth. Subscribers began using the new system to identify new mining opportunities, sunken shipwrecks and toxic metallic effluents from industrial sites. QuantumAI only marketed its product to government-approved researchers from NATO countries, and its technology was export-controlled under both US International Traffic in Arms Regulations restrictions and the Wassenaar Arrangement.

In round one, satellite imagery of the South China Sea that rendered the water transparent and revealed the location of three Chinese submarines was put online, allegedly from sources linked to QuantumAI. While the Chinese team suspected that the US government was responsible, the US team worried that whoever put this information online might similarly reveal the location of US submarines. Exactly that transpired in round two. The same website posting locational data regarding Chinese submarines expanded its coverage to include a global map with additional markers for Russian and US nuclear submarines. This posed a clear risk to strategic stability in compromising the second-strike capability that stealthy nuclear-armed submarines provide.

Adding the markers for US submarines made it clear that the revelatory website was not affiliated with the US government. Its registration history and hosting service provider in Switzerland suggested it was actually linked to the World Peace Guardians. When US Department of Homeland Security and FBI officials approached large US social-media firms for information regarding World Peace Guardians accounts, the companies replied that they would 'share relevant information with the government when it becomes available'. The US president did decide that QuantumAI hard drives could be seized, although the CEO of QuantumAI was not fully cooperative. The idea of trying to seize World Peace Guardians servers in Switzerland was discussed, but rejected.

The QuantumAI aspect of the exercise was one manifestation of what the background paper described as a 'black box model extraction' vulnerability. Such an extraction reverse-engineers an AI system to determine its parameters. An adversary may be able to use this information to enhance the effectiveness of future operations against the system by stealing intellectual property; identifying sensitive or proprietary information related to the system's training data or objectives; or developing 'adversarial inputs' to be covertly introduced into the original AI system. Such inputs confuse the system's classifiers or its pattern-recognition function, thereby causing it to miscalculate, misclassify or misinterpret elements in its operational environment.

Malfunctioning navigational systems

Malfunctioning sensors of a different kind could have contributed to major-power conflict had the game played out a little differently. In the fictional lead-up to the June crisis scenario, a US Navy destroyer conducting a freedom-of-navigation exercise in waters claimed by China in the South China Sea came within 50 metres of colliding with a People's Liberation Army Navy vessel when its automated navigational systems were experiencing serious malfunctions. While their origin was unclear, adversarial inputs may have been injected into the ship's AI system. Attacks employing them can bypass system classifiers used for cyber security, such as worm-signature generation, spam filters, distributed denial-of-service attack detection and portable-document-format malware classification. By slightly altering an image's pixels or patterns – changes that might be undetectable to the human eye – these inputs can also cause an AI system to mislabel an image.

* * *

The tabletop exercise was intended to identify potential vulnerabilities created by artificial intelligence for nuclear command and control, including strategic-warning capabilities, which could then feed into a broader policy discussion. Five lessons emerged as the basis for future policy-relevant work.

Firstly, AI cannot be fully trusted. In particular, it can be risky to rely heavily on situational awareness generated by AI outputs. In the hypothetical scenario, both the US and Russian teams discounted the AI warnings that appeared inconsistent with human observations and were later found to have been generated by false data. This raises the possibility that resilient doubts about the reliability of AI could significantly attenuate its operational and strategic impact.

Secondly, regarding emerging technology and nuclear strategy, the roll-out of new technologies – in the case at hand, sensor technology for detecting submarines – affects states differently depending on their strategic force structure. The Chinese team was most alarmed about the technology because, among the major players, it has the smallest submarine fleet and one that so far does not range far beyond China's adjacent waters. Once identified, its submarines can thus be more easily neutralised. The exposure of China's submarines before those of other players reinforced Beijing's concerns that it was being targeted.

Thirdly, shared concerns about AI-generated disinformation could foster collaboration among states to address the problem via confidence-building mechanisms. In the exercise, when the operating systems of Chinese nuclear power plants 'raised the warning noise floor' in ways that appeared intended to weaponise them, the US team shared information it had about similar noise warnings in American plants. The IAEA became the forum of choice for cooperative efforts towards a solution. More broadly, the potentially destabilising risks inherent in emerging technologies, as demonstrated in the Chinese nuclear-power example, could push states to proactively promote arms control. In our scenario, this prospect loomed only as the world stood on the brink of disaster. The takeaway is that policymakers can be educated in advance about such risks before they actually arise in a crisis and work to mitigate them.

Fourthly, the role of the private sector differs among the major powers. In the exercise, the US government had limited means of persuading the developer of the ocean-see-through sensor technology to share its data. If the company in question had been Chinese or Russian, those governments could have compelled cooperation fully and immediately. In any case, the

exercise illuminated the need for governments to appoint senior officials conversant with science and technology who either fully understand and appreciate the effects and ramifications of technological development, or cultivate regular access to experts who do.

Finally, while disinformation is a long-standing intelligence and strategic problem that pre-dates the cyber age, the integrity of AI systems is especially vulnerable to it. The combination of reliance on AI and its appropriation by malevolent actors could seriously amplify the disruptive effects of disinformation campaigns. Governments could play a key role in establishing regulatory frameworks that shape how new technology interfaces with the current suite of disinformation tools, such as social media. Although time limitations precluded further exploration of an AI system's vulnerability, hints and suspicions of AI-linked espionage permeated the tabletop exercise. They are likely to persist in the real world as well.

Acknowledgements

This article draws on background materials prepared by Sean Kanuck, then-director of the IISS Cyber, Space and Future Conflict Programme, who led the tabletop exercise and subsequently discussed with the author lessons that emerged from it.

Notes

1 See generally Kenneth Payne, 'Artificial Intelligence: A Revolution in Strategic Affairs?', *Survival*, vol. 60, no. 5, October–November 2018, pp. 7–32.