Royal United Services Institute
for Defence and Security Studies

Conference Report

Lethal Artificial Intelligence and Autonomy

Peter Roberts

Lethal Artificial Intelligence and Autonomy

Peter Roberts

RUSI Conference Report, December 2018

Royal United Services Institute
for Defence and Security Studies

187 years of independent thinking on defence and security

The Royal United Services Institute (RUSI) is the world's oldest and the UK's leading defence and security think tank. Its mission is to inform, influence and enhance public debate on a safer and more stable world. RUSI is a research-led institute, producing independent, practical and innovative analysis to address today's complex challenges.

Since its foundation in 1831, RUSI has relied on its members to support its activities. Together with revenue from research, publications and conferences, RUSI has sustained its political independence for 187 years.

Published in 2018 by the Royal United Services Institute for Defence and Security Studies.

RUSI Conference Report, December 2018

Royal United Services Institute
for Defence and Security Studies
Whitehall
London SW1A 2ET
United Kingdom
+44 (0)20 7747 2600
www.rusi.org

RUSI is a registered charity (No. 210639)

Lethal Artificial Intelligence and Autonomy

Introduction

RUSI CONVENED A CONFERENCE on Lethal Artificial Intelligence (AI) and Autonomy on 7 November 2018. The aim of the event was to illuminate some of the challenges that are unique to military forces in the use of AI and autonomy, in other words, when systems are designed to take life as opposed to those that might cause injury or death as unforeseen consequences (for example, when a driverless car kills a passer-by). It seems that few of these considerations have had a place in the deliberations and musings of Western governments, as evidenced through policy and doctrine, even when clearly related research has taken place.

The conference was designed to stimulate a discussion about the ethics, morality and legal aspects of lethal AI and autonomous systems, but solely within the context of war and warfare. Military personnel participated in the conference, along with around 90 international delegates from the academic, government, military, non-government, charity, and industrial sectors.

The event also sought to extend the debate beyond the usual narrative that focuses on 'killer robots'. It is noteworthy that UK military forces already envisage a significant role for autonomous AI in decision-making, intelligence analysis, and command-and-control functions within their force structure. US, French and German doctrine signpost similar investment and development paths. Much of this Western military orthodoxy appears to be based on the presumption of an inevitability of AI and autonomy leading to a Revolution in Military Affairs (often referred to as RMA, with the emotive historical connotations that such an expression brings to bear with military audiences). Yet, since the inception of AI, wars have become more expensive and its much-predicted revolutionary effect has failed to materialise; if anything, the lack of radical change has merely emphasised thousands of years of rather slower and more considered military evolution. AI and autonomy might therefore be more about enhancing the current paradigm of war than its revolutionary overthrow. Despite this, the weight of discussion in Western military circles continues to revolve around AI and autonomy as a transformation, without evidential support.

Lessons and Deductions

Several important themes for war and warfare emerged from the conference.

The arrival of a single, all-knowing AI – a technological singularity (the hypothesis that the invention of artificial superintelligence will abruptly trigger runaway technological growth, resulting in unfathomable changes to human civilisation) – is unlikely.[1] Indeed, the sophistication

1. Anthony Berglas, *When Computers Can Think: The Artificial Intelligence Singularity* (Createspaces, 2015).

of current AI systems (narrow AI) only exceeds human performance in specific tasks, for example in the infamous Go competition. By contrast, general or strong AI will be those systems capable of performing any task a human can and with the ability to interact with other machines. There is a small leap for such systems to become sentient, and eventually all knowing, all understanding, and – potentially – all controlling. In becoming conscious, some argue that AI will define the end of humanity. Yet the preconception and popular narrative of 'strong' AI fighting equally 'strong' AI, developing predictive and counter-predictive strikes and perfect actions appears to be shaped more by science fiction than we might care to admit. Rather, it seems that the personalities of coders become very clear within such systems and in their subsequent actions. AI systems will be a product of the differing biases, heuristics and blind-spots of programmers. Culture, education, societal upbringing and cognitive dissonance will all emerge as identifiable characteristics in differing AI systems: fingerprinting of AI decisions to states may be more possible than otherwise imagined.

The West should not be seduced by an aspiration for self-regulating AI communities (where cooperating AI systems inter-relate to regulate behaviours). Differing national ambition, potentially manifest through AI and autonomous system action, is more likely to deliver disconnected and stove-piped systems than an inter-related network of cooperative systems. Lessons from both the security and defence intelligence and cyber domains are useful in understanding this facet.

Equally, greater understanding of the data from which machine-learned 'wisdom' is being derived must be understood. Simply accepting that all data and information that is available is true would be a mistake that might lead to problematic decisions by autonomous systems and might have a disproportionate impact within AI algorithms. Enabling systems to judge the veracity of publicly available information, as well as intelligence gleaned covertly, will be a core requirement in system programming. Currently, there is more research in tailoring information feeds (risking greater confirmation bias – a 'hall of mirrors'), than in recognising challenge and 'truth'.

Programmers and coders are not useful in developing AI and autonomous systems, either for warfighting or in statecraft roles associated to war and conflict, if they do so in silos. The neurological preconditions of STEM (science, technology, engineering, and mathematics) minds create predictable, architectural processes that are good at creating code, but not in creating behaviours. Coding alone will not enable the translation of nuanced policy into computer programmes that could accurately represent changing state-level priorities, sentiment and calculus. As Google, Facebook, Decoded[2] and ZTE[3] have found, the future need is for skills that can teach systems expected normative behaviours, ethical and moral boundaries, and the intricacies of weighted proportionality in action. For example, how would you code the 'right' response to an eight-year-old child who was providing targeting information to terrorists

2. Decoded, 'About', <https://decoded.com/section/about/>, accessed 12 December 2018.
3. ZTE, 'About Us', <https://www.zte.com.cn/global/about/corporate_information/Introduction>, accessed 12 December 2018.

about to undertake an attack? In the gaming industry an important shift has taken place from systems architecture to enterprise architecture as the primary design discipline, which has been happening since 2011. Enterprise architecture attempts to design systems to support human endeavours, first by understanding how those systems will be used. Systems architecture, by contrast, tends to force humans to change behaviour to support digital systems. The integration of arts and STEM perspectives – perhaps even led by artistic minds – will be essential if AI and autonomy is to be successfully employed as a tool of war and warfare.

Combat Teaming (as espoused in the US Third Offset Strategy) is not about AI and human interface. Instead, successful teaming is more about combining AI and/or autonomy with other processes and systems. In the commercial world, the success of broader AI (for instance, that not designed for a specific task such as chess or Go), has been in applying AI with, for example, profiling to deliver targeted advertising in a more coherent manner. Understanding where and how to apply AI in warfare has yet to be considered: does AI-enabled long-range artillery support allow for more effective predictive counter-battery fire (even if this is allowed under the laws of war)?

The costs of AI and autonomous systems are not well understood, particularly in the case of security and military forces. The notion that such systems would be cheaper than similarly equipped humans lacks a sound evidence base. The conjecture over assumptions of economies in operating costs, deliverable effects and manpower savings ignore the non-financial, value-chain benefits of the alternatives. Even the presumptions over financial savings in capital outlay lack evidence against current platform and defence inflation figures. The second and third order consequences of shifting force design needs more detailed research before an evidence-based, investment appraisal decision can be made.

Physics still applies to AI and autonomous systems. The idea that UAVs, for example, might be more lethal and more survivable in air-to-air combat has some evidential support. However, the physics of getting these platforms to the location of such a fight determines platform size, aerodynamic properties and performance. When such realities are imposed on the evidence base, the outcome is less compelling.

The enemy still gets a vote. Whether all belligerents possess equal AI and autonomous capabilities or not, the restrictions imposed by moral, legal and ethical boundaries determines the outcome of conflict more than technological superiority. Adaptation, invention and innovation in combat outpaces peacetime developments markedly: and occurs for all protagonists, not simply one's own systems. Successful counters to both AI-enabled and autonomous systems already exist in terms of electronic warfare techniques, from jamming connectivity to electro-magnetic pulses that disable core processing capability. AI systems cannot be presumed to be immune from defeat, even when they have edge processing and full autonomy.

Choices

The conference concluded that there are some distinct choices for militaries to make over the next five years regarding lethal AI and autonomy. Six of the key choices highlighted at the conference are outlined below.

1. Will intelligence analysis (the basis of future autonomous decision-making) be optimised for qualitative or quantitive evidence? Different algorithms and protocols exist for each, and both have biases, heuristics and blind-spots. Combining the two approaches does not seem compatible with system architectures within 30 years.
2. Will we continue to need systems that make sub-optimal decisions rather than seeking a fixed outcome? Importantly, humans (particularly politicians) are well versed in making decisions that create opportunities of de-escalation. Shifting decision-making and recommendation processes to an autonomous system that exploits AI is not within the current research agenda, nor the production schedules of coders.
3. How much should governments be allowed to psychologically manipulate their own electorate? If adversaries and competitors will be using AI systems to subvert the sentiment of Western populations, is it ethically acceptable (perhaps even necessary) for Western state apparatus to manipulate their own electorate by, for example, tailoring what each is allowed to see? This is more complex than the obvious and well-honed question of countering confirmation bias in senior decision-makers from their own sources.
4. Is there an aspiration to have systems that serve the overall aims rather than simply importing and adapting commercial models? This plays to the question of whether the challenges faced by statecraft, war and warfare (human endeavours) have unique challenges for which AI and autonomy are distinctly unsuited.
5. What methodology might Western militaries and decision-makers use to get ahead of their adversaries and competitors? Options already exist between additional bureaucracy (as observed in the US model of building a new 'Army Futures Headquarters')[4], process (the UK model as highlighted in the MoD's Development, Concepts and Doctrine Centre's pamphlet on human-machine combat teaming)[5], education (the Chinese model of sending all senior leaders to educational establishments to specifically study AI and autonomy), or gaming (the Russian approach of developing military strategy and plans by gaming scenarios with a diverse team of multi-disciplinary actors under military tutorage).
6. If decisions by general, strong or 'super' AI can be made so fast that humans are unable to assess them – creating a 'hyperwar' scenario – does this herald a change to the fundamental nature of statecraft (political and diplomatic decision-making) in war and warfare? How content are we to hand control for actions, in such cases, to systems rather than humans?

4. Cole Stevens, 'The Army's New Futures Command Will Succeed or Fail by Congress's Hand', *The National Interest*, 29 July 2018.
5. Ministry of Defence, 'Human–Machine Teaming', Joint Concept Note 1/18, May 2018.

Conclusions

The orthodoxy in military narratives makes the march towards AI and autonomous systems within conflict, warfare, and wars more generally, seem inevitable. Not only is the West chasing the dream of lower-cost, more effective, more lethal systems in the hope of developing a competitive edge, perfect situational awareness, an understanding of intent, and perfect 'enlightened' decisions, but so are China and Russia. Yet many of the preconceptions of these discussions are not based on evidence or research. Instead the reliance on science-fiction novels and the hype generated by the actions of commercial actors has dominated the emergence of new ambition. Much of this is false and is leading to unverifiable assumptions.

The reality is that all AI systems will be the product of human biases, heuristics and blind-spots. If we are open to the idea that humans make bad decisions, we should also be open to the idea that AI systems will make bad decisions – just faster. Understanding the basics of machine learning will be a vital skill for commanders (political and military) in the future to mitigate these symptoms. The foundational programming of judgements and style, embedded by programmers, will do much to determine options presented to decision-makers. Even in such circumstances, deciding whether the human is to remain in control, or simply a figurehead to an unfulfilled and unproven hypothetical AI system is a seminal question to which this generation of leaders should be applying themselves.

Finally, it must be observed that the question of ethics and morality in lethal AI and autonomous systems differs in wars of choice as opposed to wars of necessity. Liberal values expounded by leaders in mostly peaceful times, where conflict might be over principle and values rather than those related to existential threat, apply greater restrictions to the actions of democratic states and their weapon development programmes. The logic is clear: a failure to fight within the liberal values that one espouses makes any victory valueless – even if that entails a greater cost in blood and treasure. Yet such decisions in wars of necessity are more complex and less clear. History instructs that in such circumstances, morals and ethics can become a secondary consideration to success. The question in terms of AI and autonomy is how such flexibility is embedded into systems that allows for a shift of values: should it become necessary? Indeed, is the very ability to change counterproductive to the liberal values one espouses during periods of non-war?

Core to much of the discussion over ethical, legal and moral challenges facing those charged with lethal AI and autonomy is where responsibility, culpability and attribution for decisions may lie. The current legal frameworks for war and warfare might be translated to this area without great rancour, but the ability of assurance in decision-making, of holding someone accountable and the ability to place blame is less clear. In weighing such seminal considerations, lethal AI and autonomy might perhaps herald not a revolution in military affairs, but a revolution in military ethics.

Peter Roberts is the Director of the Military Sciences research group at RUSI.