Horizon Scanning Series

The Effective and Ethical Development of Artificial Intelligence: An Opportunity to Improve Our Wellbeing

Geopolitics

This input paper was prepared by Nicholas Davis and Jean-Marc Rickli

Suggested Citation

Davis, N and Rickli, JM (2018). Geopolitics. Input paper for the Horizon Scanning Project "The Effective and Ethical Development of Artificial Intelligence: An Opportunity to Improve Our Wellbeing" on behalf of the Australian Council of Learned Academies, www.acola.org.

Submission to The Australian Council of Learned Academies and the Commonwealth Science Council on the opportunities and challenges presented by deployment of artificial intelligence

Nicholas Davis, Head of Society and Innovation, World Economic Forum
Jean-Marc Rickli, Head of Global Risk and Resilience, Geneva Centre for Security Policy

The Economics and Strategic Implications of AI

Artificial Intelligence (AI) – by which we mean artificial narrow intelligence or ANI[i] – is a general purpose technology[ii] and therefore a dual-use technology. While AI has been a field of interest and study for decades, its recent ascendance is due to increases in computing power (including the advent of specialised chips), and increases in data availability that flow from the rise of social media, the digitisation of the global economy and the development of the Internet of Things (IoT). While the term AI covers a wide variety of analytic techniques, the most recent systems can out-perform humans and form inferences from large, complex data sets to solve problems such as classification, continuous estimation, clustering, anomaly detection, recommendations, data generation and ranking.[iii] These techniques have resulted in advances in computer vision, natural language processing (NLP), robotics, planning and intelligence analysis, among others. As a result, AI systems can be used to stabilize unmanned aerial vehicles,[iv] perform automated image recognition on drone footage, and enable an automated sentry to identify humans and interpret a spoken password.[v]

The general purpose nature of AI means it can help solve problems and spur innovation in wide range of military and security related domains, including but not limited to cybersecurity, information security, diplomacy, defence, intelligence, counter-terrorism and humanitarian intervention. Recent reports published by the Center for a New American Security (Horowitz et al 2018)[vi] and the University of Oxford's Future of Humanity Institute (Brundage et al 2018)[vii] provide a useful overview of the different security domains that AI will impact, as captured in tables 1 and 2 in the Annex to this paper.

The impact of AI is often viewed from a purely technological perspective, focusing on the potential application of new capabilities. It's useful to complement this with an economic analysis of the impact of AI on security.

From an economic perspective, AI lowers the cost of two tasks that are traditionally performed by humans. First, as Agrawal et al (2018)[viii] have argued, AI drops the cost of prediction, whether that prediction is the most accurate translation of a sentence, the best next move in a video game, or if an image represents a person of interest for targeting. Consequently, and as a result of the lower cost of prediction, it drops the cost and raises the value of real-time, automated action, reaction and iterative learning.

Lower costs are thanks to rising AI productivity, scalability and open access: AI capabilities are increasing due to the combination of better algorithms, greater processing power and larger data sets[ix]; the digital nature of AI systems allows them to be distributed and scaled rapidly, allowing a single person to launch and control numerous instances; and AI algorithms are often publically published or open-source.[x] Furthermore, removing the human from the battlefield lowers risk-related human costs and can enable the greater performance of vehicles.[xi]

AI empowering new and existing actors

These economic factors incentivize new actors to explore opportunities to exploit security for economic or political gain, while the automation of tasks mean that existing actors will potentially become more dangerous. As Brundage et al (2016) have pointed out, as greater numbers of actors invest in AI-driven tactics, higher rates of experimentation and innovation will result in the emergence and proliferation of new threats and tactics.[xii] These elements, combined with the growing ecosystem of digitally-connected individuals and devices (many of which represent critical health, transport, energy and food infrastructure) mean that far greater numbers of individuals, organisations and countries become exposed to AI-enabled security threats.

Relative to other military and security technologies, the low cost, high access and broad range of capabilities of AI empowers individuals, small groups, criminal enterprises and other non-state actors.[xiii]

Particularly in the digital domain, acquiring new cyber capabilities is cheap, and the marginal cost of additional production – adding a target – is almost zero.[xiv] The advantage of AI systems in such contexts rely on the ability to enhance the productivity of human operators through creating and scaling sophisticated attacks that are difficult to defend with a corresponding defensive AI, such as automated malware and highly-personalised social engineering attacks.[xv]

In the physical domain, AI-enabled commercial products can be repurposed for surveillance or offensive uses, just as other commercial technologies are repurposed for both offense and defence by non-state actors. For example, ISIS mounts high-definition cameras under drones to improve intelligence and acquire situational awareness. They also used drones to drop 40mm grenades on Iraqi positions, allegedly killing up to 30 Iraqi soldiers per week during the battle of Mosul in 2017.[xvi] Although not relying on AI, such examples demonstrate how agile terrorist organisations are in using commercial technologies to support their goals.

This is not to imply that non-state actors will ever have an absolute advantage over state actors simply due to AI, or in the use of AI itself. In reality, as discussed below, state actors have important advantages, including access to large data sets that are difficult for non-state actors to acquire and a access to large pools of skilled professionals that can conduct research and development. However the availability of pre-trained algorithms and access to public datasets will encourage and democratise experimentation in offensive strategies by smaller, less well-resourced groups. In this way, diverse, malicious applications that leverage AI as an enabling technology may reshape strategic environments on par with the use of IEDs in Iraq and Afghanistan, with similar potential implications for investment in defensive strategies.[xvii]

Training with and learning from big data as a critical enabler of AI efficacy

An essential input for the development of AI systems is access to suitable training data. While some of the most sophisticated algorithms are publically available and suitable processing power can be rented cheaply through the cloud, many AI techniques require large, labelled datasets to produce accurate results. While there are a range of public datasets available for machine learning,[xviii] the most valuable, and useful data is closely held by the companies or states that gather it.

One implication of the importance of large, high-quality datasets in AI is that non-democratic countries and regimes, with the ability to leverage state-wide access to data and conduct population-level experiments, will have an advantage in the development and deployment of AI systems.

This advantage is already evident in global AI competitions. In Nov 2017, Yitu Tech, a Shanghai-based AI start-up, won the IARPA Face Recognition Prize Challenge. As the company's press release put it, "China is quietly leading the world in deploying AI in real-world applications." [xix] The FT reports that Yitu's face recognition system was built for Chinese law enforcement,[xx] and, according to their corporate brochure, is "capable of identifying over 1.8 billion individuals within seconds".[xxi] China aims to be the leader in AI and cyber technologies in the next decade, with security as well as economic progress in mind.[xxii] In addition to its own use of AI-enabled citizen surveillance and "the gamification of obedience", Chinese AI surveillance technologies are being exported to other states such as Zimbabwe, Singapore, Malaysia or Mongolia.[xxiii]

Given the role that data plays at the frontier of AI applications, the so-called "AI arms race" will therefore be influenced by a combination of private and public sector investment, willing talent to develop applications and access to suitable datasets related to individuals, objects and the natural world.

The implications of uncertain, opaque offensive capabilities

The early-stage nature of many AI applications, the hype surrounding AI in general and the potential for AI to enable new forms of offensive strategies could destabilize security relationships.

As discussed above, AI's use in the cyber domain benefits from the asymmetric cost of engineering new attacks versus developing counter-defensive measures, providing an advantage to the attackers.[xxiv] In the physical domain, AI-enabled approaches such as swarming and the use of lethal autonomous weapons systems (LAWS) may reduce the supremacy of defence.

Swarming relies on overwhelming and saturating the adversary's defence system by synchronizing a series of simultaneous and concentrated attacks.[xxv] Swarming approaches rely heavily on AI, and related algorithms and physical systems are maturing rapidly. In October 2016, the US Department of Defense conducted an experiment that saw 103 Perdrix micro drones autonomously deal with four different objectives. Meanwhile, the world record for swarming drones was broken by a Chinese company, EHang, in May 2018 with an AI-assisted swarm of 1374 drones flying over the City wall of Xi'An.[xxvi] While they are able to be used for both defence and offense, the development of LAWS that could be used in swarms could have a profoundly destabilising impact on strategic stability.[xxvii] At the extremes, if the application of AI-driven tactics promises to reduce the effectiveness of second strike retaliatory capabilities (as posited in P W Singer and August Cole's 2015 novel *Ghost Fleet*),[xxviii] it follows that deterrence will be replaced by pre-emption.[xxix] This encourages escalation and arms races.[xxx]

A contributing factor to this is that the promise of AI may lead countries to over-estimate its abilities to deal a 'knock-out blow', and initiate conflict accordingly. Lawrence Freedman points out in *The Future of War: A History* that the history of warfare features a regular assumption that "the odds of success might be shifted decisively as a result of some new technology".[xxxi] The danger is that AI may encourage "a fantasy of war that [is] fast, easy and decisive", a fallacy that, in H. R. McMaster's words, fails to recognise the "uncertainty of war, the trajectory of which is constantly altered by varied interactions with determined and elusive enemies."[xxxii] Building on these ideas, the emergence of AI systems is strategically influential because their true contribution to strategic or tactical effectiveness is highly uncertain at their current stage of development: overconfidence in one's own abilities conveyed by AI encourages initiation or escalation and the overestimation of the offensive abilities of others encourages pre-emption.

Further complicating this picture is the opaqueness and brittleness of many AI systems, particularly those that rely on deep learning methods. Verifying AI algorithms, assuring their reliability under a

wide set of conditions and combating (currently inherent) vulnerabilities (such as the ability to 'spoof' image recognition systems) is extremely challenging.[xxxiii] In addition there is the danger that AI systems themselves may escalate a security threat or make flawed decision making that compromises an entire organisation through automated decision making – a scenario that has played out in financial markets in the 2010 flash crash[xxxiv] and in the 2012 bankruptcy of Knight Capital Group.[xxxv]

Governing AI with security in mind:

Governing AI with its security implications in mind will require stakeholders in Australia and New Zealand to consider at least three strategies.

First, it may mean rethinking the openness with which new AI techniques are developed and shared, and how to make better use of large data sets whilst protecting the rights of citizens. This may include creating new ways to manage access to "public good" data sets and algorithms that support legitimate research while discouraging training of systems that directly threaten human security.

Take for example the national and regional projects drawing on the Open Data Cube project, an initiative that has emerged from Geoscience Australia, to increase the value and impact of global Earth observation satellite data by providing an open and freely accessible exploitation platform.[xxxvi] At low to medium resolutions in the Australian context, such a platform for developing and running advanced algorithms is uncontroversial and extremely positive, enabling innovative solutions in the environmental, agriculture, urban planning and resources spaces. However at higher resolutions – and higher refresh rates – in the context of other continents where border disputes, refugee flows and illegal activity proliferate, the ability for both state and non-state actors to use such a platform to track and target groups of people or natural resources is of significant security and ethical concern. These challenges have provoked the creation of new norms and ethical frameworks around the use of data in sensitive and security-related contexts where humanitarian law applies, such as *The Signal Code*.[xxxvii]

Recent examples with commercial fitness tracker applications such as Strava or Polar have raised similar major security concerns for law enforcements and military personal by revealing their positions in operational duty but also at home.[xxxviii] Rethinking access to "public good" data sets is not only restricted to AI research and applications. A paper explaining the synthesis of horsepox was published earlier this year.[xxxix] This led to a controversy in the scientific community and raised some alarms as it might give terrorists or other malicious actors a recipe to synthesise smallpox virus that was eradicated by the international community in 1980.[xl] A general reflections on the security implications of the applications of the Fourth Industrial Revolution (4IR) technologies has to take place at the global level as their disruptive impact could be very destabilizing.

Second, governance will require far greater international cooperation, dialogue and rule-development. This is particularly required for discussions and the creations of new norms around autonomous weapons systems.

The United Nations through the Convention of Certain Conventional Weapons has debated the issue of LAWS since 2014. As yet, no consensus has emerged on a definition of these weapons, let alone whether or how they could be banned or limited. While the group of government experts involved in the negotiations do agree that meaningful human control must be retained in the use of these weapons, the precise definition of meaningful human control and the practicalities of implementing such requirements remain outstanding. This will likely remain a major bone of contention as states have now very different interests regarding the militarisation of artificial intelligence. As Vladimir Putin stated, "the one who becomes the leader in this sphere [artificial intelligence] will be the ruler of the world."[xli] Given that autonomy is being seen as the new 'silver bullet' of the 21st Century, it is therefore highly doubtful that any major treaty banning the development of LAWS will ever be concluded.[xlii] Due

to the 'black box' nature of many machine learning algorithms that makes tracing of their decision-making very difficult, if not impossible, it is non-state actors or military organisations that value deniability over predictability that will most likely be early adopters of autonomous systems.[xliii]

Governments and international organisations are becoming more aware of the threats related to AI. On 12 July 2018, the UN Secretary General appointed a High Level Panel on Digital Cooperation. One of the goal of this panel is to mitigate the risks and curtail any unintended consequences of digital technologies. The panel's ambition to support "cooperative and interdisciplinary approaches to ensure a safe inclusive digital future for all taking into account relevant human rights norms" is a step in the right direction,[xliv] recognising that, given the rapid pace of development of emerging technologies, traditional governance systems are failing to serve their purpose.

Third, new consortia of actors must be closely involved in innovative governance efforts, particularly the private sector. National and intergovernmental efforts will not be sufficient to govern the security threats posed by AI. The vast majority of AI research and application – and funding for both – is occurring in the private sector, meaning that the private sector is a more influential actor in the area of AI than in relation to other security-relevant technologies.[xlv]

Indeed, when it comes to engaging with challenging questions around the governance of AI, the private sector is moving more boldly than most public sector bodies. Examples include the Future of Life Institute, which gained particularly high visibility in 2015 for issuing an Open Letter that gathered over 8'000 signatures, on *Research Priorities for Robust and Beneficial Artificial Intelligence*. The letter called for verification measures, security against unauthorized manipulation, and methods for continuous and reliable human control of AI as important areas of research.[xlvi] The Partnership on AI, a non-governmental organization founded by a coalition of tech giants: Amazon, Google, Facebook, IBM, Microsoft and Google, aims to raise awareness on AI technologies and to develop and share best practices in the research, development and fielding of AI technologies.[xlvii] OpenAI, a non-profit AI research company sponsored by individuals such as Peter Thiel or Elon Musk and companies such as Microsoft and Amazon, seeks to build safe artificial general intelligence and ensure that AGI's benefits will be as widely and as evenly distributed as possible.[xlviii] The World Economic Forum's Center for the Fourth Industrial Revolution in San Francisco, brings together multistakeholder groups of leaders and experts to create technology policy and governance models across eight emerging technologies, including AI.[xlix]

Meanwhile, the employees of companies developing AI systems are influencing the debate on their use in security contexts. For example, in April 2017 the US Department of Defense partnered with Google to automate image recognition in real time for drone footage analysis. Project Maven reached promising results with more than 80 percent identification accuracy until, in April 2018, 3000 Google employees signed an open letter to Google CEO Sundar Pichai asking the company to terminate its Project Maven contract and commit that Google nor its contractors will ever build warfare technology.[l] On 1 June 2018, Google announced that it would not renew the contract with the US DoD that is due to expire in 2019[li] and on 7 June 2018 launched a new set of principles governing their development and use of AI.[lii] Similar initiatives aiming at restricting or cancelling cooperation with law enforcement authorities and militaries were initiated by employees of Western tech companies such as Amazon, Salesforce and Microsoft.[liii]

Currently, leaders in the tech industry and the scientific community play the most active roles in awareness raising and cooperation on AI. As governance concerns begin to crystallize into critical risks, it is essential for Australian leaders to increase the engagement of a large range of actors from private and start-up companies to governments, international organizations and the academic community, to collaboratively develop a set of approaches and safeguards for a safe future in the context of AI.[liv]

Annex

1. Defining Artificial Intelligence

A common, though incomplete, definition of Artificial Intelligence (AI) is "the capability of a computer system to perform tasks that normally require human intelligence."[lv] It is important to distinguish between three different conceptions of AI, which generate different sets of dynamics and concerns.

The first class of AI is the most in line with the common understanding of the term "artificial intelligence" in referring a system which can perform tasks normally expected of a human, better known as artificial general intelligence (AGI). While it does not yet exist, AGI is (theoretically) able to operate in much broader and less certain contexts than ANI, mimicking the broad cognitive flexibility of the human brain. In a recent survey of AI experts, the median timeframe predicted for the achievement of AGI is 45 years from now.[lvi]

The second and most distant class of AI is artificial super intelligence (ASI). This form of AI describes a general intelligence that massively exceeds the capabilities of the human brain. As Nick Bostrom[lvii] and others have detailed, it is fair to assume that ASI would potentially represent an existential threat to humanity, creating a set of security risks far beyond the range of existing, human-driven concerns.[lviii]

The third and most basic class of AI is artificial narrow intelligence (ANI), enabled by a wide variety of machine learning approaches. Machine learning consists of the development of algorithms which progressively improve performance on a specific task by making and testing predictions on data without being explicitly programmed. ANI examples include the algorithm behind Google Translate, spam-filtering systems, facial recognition technology and algorithms designed to learn and play video games etc. Many ANI systems can now outperform human beings at specific tasks, including gaming[lix] and image recognition. However, the brittleness of ANI systems means that performance drops dramatically as conditions change or inputs become distorted.[lx] Today, only ANI exists, and hence this paper deals exclusively with this class of AI.

2. Mapping potential applications of AI to security domains

Table 1: Evolving offensive and defensive strategy by security domain, based on Horowitz et al 2018

Domain	*Offense*	*Defence*
Cybersecurity	Automation which lowers the bar for attacks by individuals and non-state groups and increases the scale of potential attacks for all actors	Automated Red-teaming and Software Verification and Validation
	Exploring New Cyber Vulnerabilities and Attack Vectors	
	Automated Customized Social Engineering Attacks	
Information Security	Targeted Propaganda and Deep Fakes, exploiting behavioural data, amplification and agenda setting, sentiment targeting,	Countering Disinformation via automated vetting, fake news detection, trollbot blocking, verification of authenticity
Economic and Financial Tools of Statecraft	Financial market manipulation and disruption	Strengthening counter–illicit–financing operations
Defense	Situational awareness	
	Electromagnetic spectrum dominance	
	Decoys and camouflage	
	New tactic generation via simulated environments	

	Autonomous command and control systems	
	Swarming attacks	Counter-swarming attacks
Intelligence	Trend analysis and pattern recognition across multiple data sets	
	Profiling	Facial recognition and gait analysis
	Counter-AI spoofing	
	Back story generation	
	Automated intelligence analysis	
Diplomacy and Humanitarian Missions	Electorate manipulation	Election monitoring
		Logistics and planning support

Source: Adapted from Horowitz et al (2018)

Table 2: Evolving security threats and examples, based on Brundage et al 2018

Security domain	*Threat*	*Examples*
Digital security	Automation of social engineering attacks	Customised malicious websites/emails/links targeting an individual; chatbots masquerading as a contact
	Automation of vulnerability discovery	Historical patterns of code vulnerabilities are used to speed up the discovery of new vulnerabilities, and the creation of code for exploiting them
	More sophisticated automation of hacking	improve target selection and prioritization, evade detection, and creatively respond to changes in the target's behaviour
	Human-like denial-of-service	A massive crowd of autonomous agents overwhelms an online service by imitating human behaviour in click patterns and navigation, preventing access from legitimate users and potentially driving the target system into a less secure state
	Automation of service tasks in criminal cyber-offense.	Automation of complementary tasks that make up their attack pipeline, such as payment processing or dialogue with ransomware victims
	Prioritising targets for cyber attacks using machine learning	Large datasets are used to identify victims more efficiently
	Exploiting AI used in applications, especially in information security.	Data poisoning attacks that reduce AI system effectiveness or are used to gain access to a system
	Black-box model extraction of proprietary AI system capabilities	
Physical security	Terrorist repurposing of commercial AI systems	Commercial systems are used in harmful and unintended ways, such as using drones or autonomous vehicles to deliver explosives and cause crashes.
	Endowing low-skill individuals with previously high-skill attack capabilities	AI-enabled automation of high-skill capabilities — such as self-aiming, long-range sniper rifles
	Increased scale of attacks	A single person launching an attack with many weaponized autonomous drones
	Swarming attacks	Distributed networks of autonomous robotic systems, cooperating at machine speed, provide ubiquitous surveillance to monitor large areas and groups and execute rapid, coordinated attacks

	Attacks further removed in time and space	Physical attacks are further removed from the actor initiating the attack as a result of autonomous operation, including in environments where remote communication with the system is not possible.
Political Security	State use of automated surveillance platforms to suppress dissent	automating image and audio processing, permitting the collection, processing, and exploitation of intelligence information at massive scales for myriad purposes, including the suppression of debate
	Fake news reports with realistic fabricated video and audio	Fake news reports with realistic fabricated video and audio
	Automated, hyper-personalised disinformation campaigns	Individuals are targeted in swing districts with personalised messages in order to affect their voting behaviour
	Automating influence campaigns	key influencers, who can then be approached with (malicious) offers or targeted with disinformation
	Denial-of-information attacks	Bot-driven, large-scale information generation attacks are leveraged to swamp information channels with noise
	Manipulation of information availability	Media platforms' content curation algorithms are used to drive users towards or away from certain content

Source: Adapted from Brundage et al (2018)

End Notes

[i] See Annex for definition

[ii] Drawing on the framework set up by Jovanovic and Rousseau, AI can be characterised as general purpose because of its ability to spread to most sectors, to improve over time and hence lower the costs of its users, and to be innovation spanning – to make it easier to invent and produce new products or processes. Jovanovic, Boyan; Rousseau, Peter L, (2005) "General Purpose Technologies", NBER Working Paper Series; Cambridge, Jan 2005. DOI:10.3386/w11093

[iii] Mckinsey (2018). "Notes from the AI Frontier: Insights from Hundreds of Use Cases", Discussion Paper, available at: https://www.mckinsey.com/~/media/mckinsey/featured%20insights/artificial%20intelligence/notes%20from%20the%20ai%20frontier%20applications%20and%20value%20of%20deep%20learning/mgi_notes-from-ai-frontier_discussion-paper.ashx

[iv] Prateek Burman (2016), "Quadcopter stabilization with neural network", Masters Thesis, Department Of Electrical and Computer Engineering, University of Texas at Austin, available at: https://repositories.lib.utexas.edu/handle/2152/45875

[v] Pike, John (2011). "The Samsung Techwin SGR-A1 Sentry Guard Robot". Global Security, 7 November2011, available at: https://www.globalsecurity.org/military/world/rok/sgr-a1.htm

[vi] Michael Horowitz, Paul Scharre, Gregory C. Allen, Kara Frederick, Anthony Cho and Edoardo Saravalle (2018), "Artificial Intelligence and International Security", Center for a New American Security, available at: https://www.cnas.org/publications/reports/artificial-intelligence-and-international-security

[vii] Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, Hyrum Anderson, Heather Roff, Gregory C. Allen, Jacob Steinhardt, Carrick Flynn, Seán Ó hÉigeartaigh, Simon Beard, Haydn Belfield, Sebastian Farquhar, Clare Lyle, Rebecca Crootof, Owain Evans, Michael Page, Joanna Bryson, Roman Yampolskiy, Dario Amodei (2018), "The Malicious

Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation", arXiv:1802.07228, available at: https://arxiv.org/abs/1802.07228

[viii] Ajay Agrawal, Gans, Joshua, Goldfarb, Avi (2018), *Prediction Machines: The Simple Economics of Artificial Intelligence*, Harvard Business Review

[ix] A recent study by OpenAI shows that while the Moore's arguing that computing power doubles every 18 months, the amount of compute used in the largest AI training runs has a 3.5 month-doubling period. While the former increased by a factor of 12 since 2012, the latter increased by a factor of 300'000. OpenAI (2018). "AI and Compute", OpenAI, 16 May, https://blog.openai.com/ai-and-compute/

[x] Patrick Shafto (2016). "Why big tech companies are open-sourcing their AI system", *The Conversation*, 22 February 2016, available at: https://theconversation.com/why-big-tech-companies-are-open-sourcing-their-ai-systems-54437

[xi] Paul Schaare (2017). "Artificial Intelligence and the Future of War", address to the U.S. Army War College, available at: https://www.cnas.org/publications/video/the-ai-revolution-paul-scharre-the-director-future-of-warfare-initiative

[xii] A similar development can be observed in synthetic biology.

[xiii] Jean-Marc Rickli (2018). "The Economic and Security Implications of Artificial Intelligence for the Arab Gulf Countries", *EDA Insights*, Abu Dhabi, Emirates Diplomatic Academy, forthcoming

[xiv] Gregory C. Allen and Taniel Chan (2017), "Artificial Intelligence and National Security," Belfer Center for Science and International Affairs, July 2017, 18, available at: https://www.belfercenter.org/sites/default/files/files/publication/AI%20NatSec%20-%20final.pdf.

[xv] Brundage et al (2018)

[xvi] Pablo Chovil. "Air Superiority under 2000 Feet : Lessons from Waging Drone Warfare against ISIL", *War on the Rock*, 11 May 2018, https://warontherocks.com/2018/05/air-superiority-under-2000-feet-lessons-from-waging-drone-warfare-against-isil/

[xvii] Jason Shell (2017), "How The IED Won: Dispelling The Myth Of Tactical Success And Innovation", available at https://warontherocks.com/2017/05/how-the-ied-won-dispelling-the-myth-of-tactical-success-and-innovation/

[xviii] See for example the GitHub repository: https://github.com/awesomedata/awesome-public-datasets

[xix] Yitu Tech (2017). "Yitu Tech Wins the 1st Place in Identification Accuracy In Face Recognition Prize Challenge 2017", PR Newswire, 3 November 2017, available at: https://www.prnewswire.com/news-releases/yitu-tech-wins-the-1st-place-in-identification-accuracy-in-face-recognition-prize-challenge-2017-300549292.html

[xx] Louise Lucas and Richard Waters (2018), China and US compete to dominate big data, Financial Times, 1 May 2018, available at: https://www.ft.com/content/e33a6994-447e-11e8-93cf-67ac3a6482fd

[xxi] Yitu (2018), Corporate Brochure, available at: http://yitutech.sg/wp-content/uploads/2018/01/YITU_Corporate-Brochure_Indiv-pages.pdf

[xxii] Cate Cadel and Adam Jourdan (2017). "China Aims to Become World Leaders in AI by 2025, Challenges U.S. Dominance," *Reuters*, 20 July, https://www.reuters.com/article/us-china-ai/china-aims-to-become-world-leader-in-ai-challenges-u-s-dominance-idUSKBN1A5103; Tencent Technology (2017). "China raises artificial intelligence to national strategy and reaches world leading level in 2030", 20 July 2017, http://tech.qq.com/a/20170720/045464.htm (in Mandarin);

[xxiii] Jean-Marc Rickli quoted in Ty Joplin (2018). "Long Form: China's Global Surveillance-Industrial Complex," Albawaba News, 21 June, https://www.albawaba.com/news/long-form-china's-global-surveillance-industrial-complex-1141152

[xxiv] Stanislav Abaimov and Paul Ingram (2017). "Hacking UK Trident: a Growing Threat," *British American Security Information Council (BASIC)*, June.

[xxv] Paul Scharre P *Robotics on the Battlefield Part II: The Coming Swarm*, Washington, Center for a New American Security, October 2014

[xxvi] EHang (2018). „EHang Egret's 1374 Drones Dancing over the City Wall of Xi'an, Achieving a Guinness World Title," accessed 21 July 2018, http://www.ehang.com/news/365.html

[xxvii] Jürgen Altman and Frank Sauer. Autonomous Weapons Systems and Strategic Stability. *Survival,* Vol. 59, issue 5, 2017, pp.117-142.

[xxviii] P W Singer and August Cole (2015). *Ghost Fleet: A Novel of the Next World War*, Mariner Books

[xxix] International law prohibits the use of force except in case of self-defense and if the UN Security Council allows it under chapter VII of the UN Charter. Pre-emption is therefore in direct contradiction to the spirit of the UN Charter and its application a violation of Art 2(4) of the Charter. See Jean-Marc Rickli (2018). "The Impact of Autonomy and Artificial Intelligence on Strategic Stability", UN Chronicle, July-August.

[xxx] Jean-Marc Rickli. "The Impact of Autonomous Weapons Systems on International Security and Strategic Stability," in Ladetto, Quentin, *Defence Future Technologies: What We See on the Horizon*. Thun: Armasuisse, 2017, pp. 61-64. https://deftech.ch/What-We-See-On-The-Horizon/armasuisseW%2BT_Defence-Future-Technologies-What-We-See-On-The-Horizon-2017_HD.pdf

[xxxi] Lawrence Freedman (2017), *The Future of War: A History*, Hachette, p 278

[xxxii] H R McMaster (2014). "Discussing the Continuities of War and the Future of Warfare", Small Wars Journal, available at: http://smallwarsjournal.com/jrnl/art/discussing-the-continuities-of-war-and-the-future-of-warfare-the-defense-entrepreneurs-foru

[xxxiii] Ben Barry (2018). "Is assurance the Achilles heel of military artificial intelligence?" IISS, available at: https://www.iiss.org/blogs/military-balance/2018/07/assurance-problem-military-artificial-intelligence

[xxxiv] Jill Treanor (2015). "The 2010 'flash crash': how it unfolded", The Guardian, 22 April 2015, available at: https://www.theguardian.com/business/2015/apr/22/2010-flash-crash-new-york-stock-exchange-unfolded

[xxxv] Nathaniel Popper (2012). "Knight Capital Says Trading Glitch Cost It $440 Million", Dealbook, 2 August 2012, New York Times, available at: https://dealbook.nytimes.com/2012/08/02/knight-capital-says-trading-mishap-cost-it-440-million/

[xxxvi] See https://www.opendatacube.org/ for more information

[xxxvii] Faine Greenwood, Caitlin Howard, Danielle Escudero Poole, Nathaniel A. Raymond and Daniel P. Scarnecchia (2017), *The Signal Code*, Harvard Humanitarian Initiative, Standards and Ethics Series 2, available at : https://signalcodeorg.files.wordpress.com/2017/01/signalcode_final7.pdf

[xxxviii] Posma, Foeke (2018). "After Strava, Polar is Revealing the Homes of Soldiers and Spies," Bellingcat, 8 July, https://www.bellingcat.com/resources/articles/2018/07/08/strava-polar-revealing-homes-soldiers-spies/

[xxxix] Noyce, Rayan and al (2018). "Construction of an Infectious Horsepox Virus Vaccine from Chemically Synthesized DNA fragments," *PLOS One*, 19 January, http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0188453

[xl] Gregory Lewis (2018). "Horsepox Synthesis : a Case of the Unilateralist's Curse?," *Bulletin of the Atomic Scientists*, 19 February, https://thebulletin.org/2018/02/horsepox-synthesis-a-case-of-the-unilateralists-curse/

[xli] Associated Press (2017). "Putin : Leader in Artificial Intelligence Will Rule the World," *CNBC*, 4 September, https://www.cnbc.com/2017/09/04/putin-leader-in-artificial-intelligence-will-rule-world.html

[xlii] Krieg, Andreas and Rickli, Jean-Marc (2019). *Surrogate Warfare : The Transformation of War in the Twenty-first Century*. Washington: Georgetown University Press (forthcoming)

[xliii] Michael C. Horowitz (2018). "Artificial Intelligence, International Competition, and the Balance of Power," *Texas National Security Review*, Vol. 1, Issue 3, https://tnsr.org/2018/05/artificial-intelligence-international-competition-and-the-balance-of-power/ and Michael Shoebridge (2018). "AI and National Security: Lethal Robots or Better Logistics," *The Strategist, ASPI*, 20 July, https://www.aspistrategist.org.au/ai-and-national-security-lethal-robots-or-better-logistics/

[xliv] United Nations (2018). "Secretary-General Appoints High-Level Panel on Digital Cooperation," Press release, SG/A/1817, 12 July, https://www.un.org/press/en/2018/sga1817.doc.htm

[xlv] There is little in the way of authoritative, global comparisons of public v.s. private sector funding for AI programmes, but current global estimates of private sector funding are in the range of US$20-30bn, while current government funding across the major economies for AI-related activities is a fraction of this when country-by-country comparisons are made. See for example McKinsey (2017). "Artificial intelligence: The next digital frontier", https://www.mckinsey.com/~/media/McKinsey/Industries/Advanced%20Electronics/Our%20Insights/How%20artificial%20intelligence%20can%20deliver%20real%20value%20to%20companies/MGI-Artificial-Intelligence-Discussion-paper.ashx ; AI Impacts (2017). "Funding of AI Research", https://aiimpacts.org/funding-of-ai-research/ ; US National Science and Technology Council (2016). "The National Artificial Intelligence Research and Development Strategic Plan", October 2016, https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf ; and Tim Dutton (2018). "An Overview of National AI Strategies", Medium.com, 28 June 2018, https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd

[xlvi] Stuart Russell, D. D. (2015). "Research Priorities for Robust and Beneficial Artificial Intelligence." *AI Magazine,* 36(4), pp. 105-114. Retrieved from https://futureoflife.org/data/documents/research_priorities.pdf?x33688

[xlvii] See https://www.partnershiponai.org/ for more information

[xlviii] See https://openai.com/ for more information

[xlix] See https://www.weforum.org/centre-for-the-fourth-industrial-revolution for more information

[l] Letter to Sundar Pichai, available at: https://static01.nyt.com/files/2018/technology/googleletter.pdf

[li] Kate Conger (2018). "Google Plans Not to Renew Its Contract for Project Maven, a Controversial Pentagon Drone AI Imaging Program", Gizmodo, 1 June 2018, available at https://gizmodo.com/google-plans-not-to-renew-its-contract-for-project-mave-1826488620

[lii] Sundar Pichai (2018). "AI at Google: our principles", The Keyword, 7 June 2018, Google, available at: https://blog.google/technology/ai/ai-principles/

[liii] Nitasha Tiku (2018). "Why Tech Workers Dissent is Going Viral,"*Wired*, 29 June, https://www.wired.com/story/why-tech-worker-dissent-is-going-viral/

[liv] Jean-Marc Rickli (2018). "International Governance and the Malicious Uses of Artificial Intelligence," *Swissfuture: Magazin für Zukunftsmonitoring*, issue 2, (forthcoming).

[lv] M. L. Cummings (2017). Artificial Intelligence and the Future of Warfare

[lvi] Katja, Grace and *al.*. "When Will AI Exceed Human Performance? Evidence from Experts." arXiv:1705.08807, 24 May 2017, https://arxiv.org/pdf/1705.08807.pdf

[lvii] Nick Bostrom (2014). *Superintelligence: Paths, Dangers, Strategies*, OUP Oxford

[lviii] WEF. *The Global Risks Report 2017.* Geneva: The World Economic Forum, 12th Edition, 2017, http://www3.weforum.org/docs/GRR17_Report_web.pdf

[lix] https://www.theverge.com/2018/6/25/17492918/openai-dota-2-bot-ai-five-5v5-matches

[lx] Samuel Dodge and Lina Karam (2017), "A Study and Comparison of Human and Deep Learning Recognition Performance Under Visual Distortions", arXiv:1705.02498v1 [cs.CV] 6 May 2017