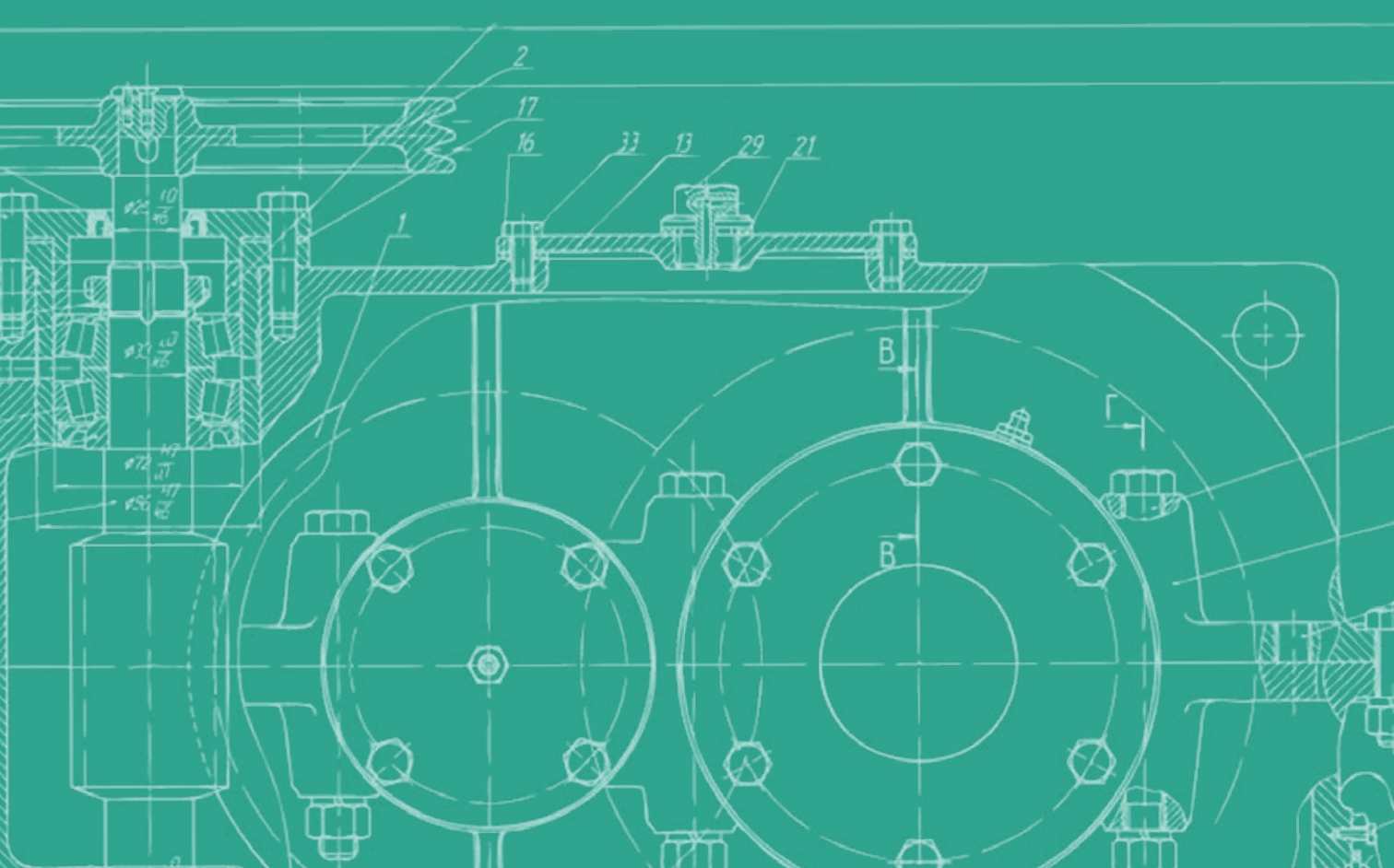
Centre for International Governance Innovation

A CIGI Essay Series

MODERN CONFLICT AND ARTIFICIAL INTELLIGENCE



CONTENTS

Credits

Managing Director & General Counsel
AARON SHULL

Managing Editor
ALLISON LEONARD

Senior Publications Editor
JENNIFER GOYDER

Publications Editor
LYNN SCHELLENBERG

Graphic Designers
SAMI CHOUHDARY
ABHILASHA DEWAN
BROOKLYNN SCHWARTZ

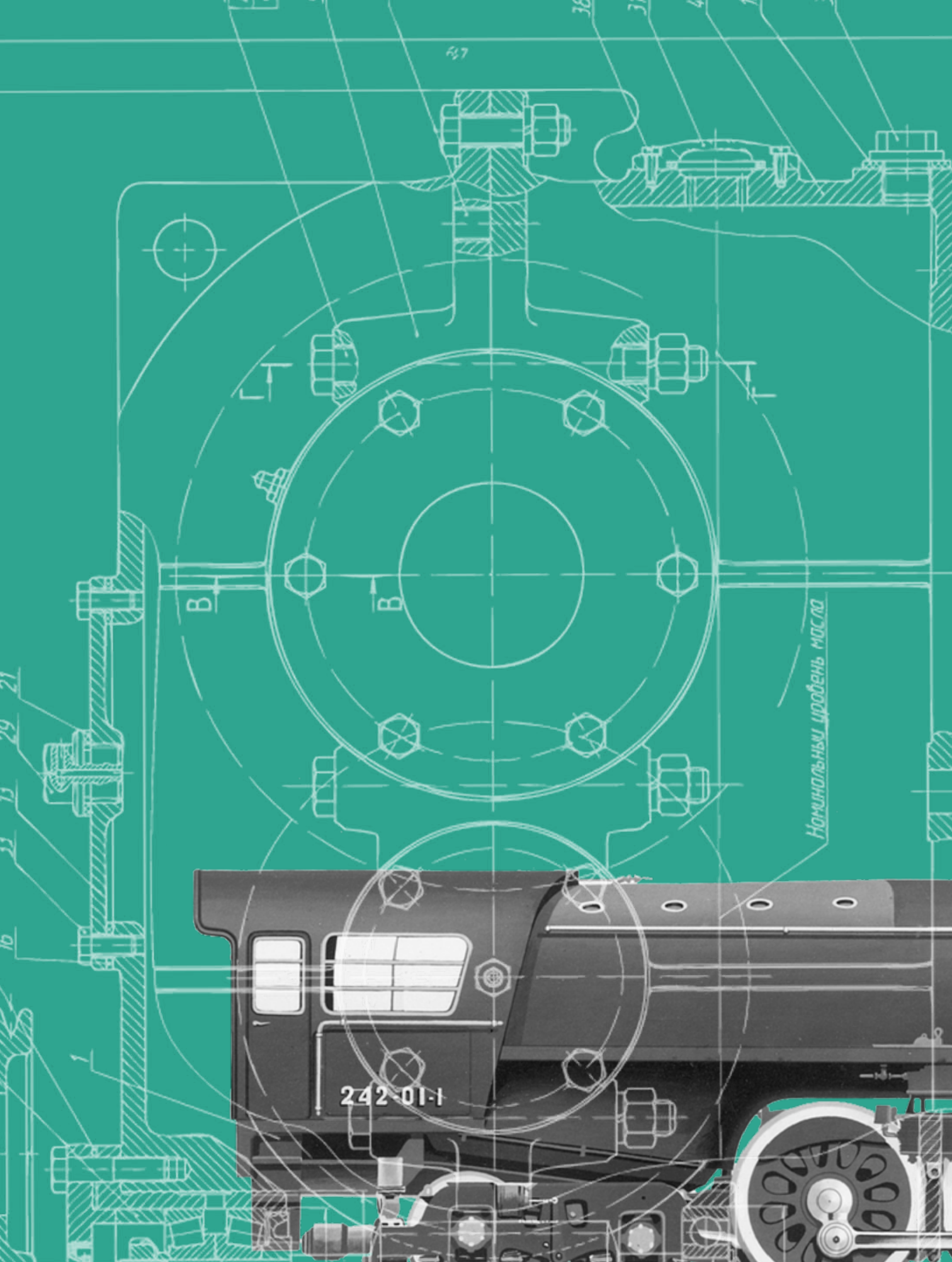
Watch the series videos at
cigionline.org/conflict-ai

Inquiries may be directed to communications@cigionline.org

Printed in Canada on paper containing 30% post-consumer fibre and certified by the Forest Stewardship Council®.

Centre for International Governance Innovation

67 Erb Street West
Waterloo, ON, Canada N2L 6C2
www.cigionline.org


242-01-1
Номинальный уровень масла

Introduction: How Can Policy Makers Predict the Unpredictable?

Meg King and Aaron Shull

Policy makers around the world are leaning on historical analogies to try to predict how artificial intelligence, or AI — which, ironically, is itself a prediction technology — will develop. They are searching for clues to inform and create appropriate policies to help foster innovation while addressing possible security risks. Much in the way that electrical power completely changed our world more than a century ago — transforming every industry from transportation to health care to manufacturing — AI's power could effect similar, if not even greater, disruption.

Whether it is the "next electricity" or not, one fact all can agree on is that AI is not a thing in itself. Most authors contributing to this essay series focus on the concept that AI is a general-purpose technology — or GPT — that will enable many applications across a variety of sectors. While AI applications are expected to have a significantly positive impact on our lives, those same applications will also likely be abused or manipulated by bad actors. Setting rules at both the national and the international level — in careful consultation with industry — will be crucial for ensuring that AI offers new capabilities and efficiencies *safely*.

Situating this discussion, though, requires a look back, in order to determine where we may be going. While AI is not new — Marvin Minsky developed what is widely believed to be the first neural network learning machine in the early 1950s — its scale, scope, speed of adoption and potential use cases today highlight a number of new challenges. There are now many ominous signs pointing to extreme danger should AI be deployed in an unchecked manner, particularly in military applications, as well as worrying trends in the commercial context related to potential discrimination, undermining of privacy, and upended traditional employment structures and economic models.

Whether it is the "next electricity" or not, one fact all can agree upon is that AI is not a thing in itself.

From a technological perspective, the drivers of the change are twofold. First is the advancement in the two methodologies employed to create algorithms: deep learning and machine learning. Machine learning, in essence, is "technology that allows systems to learn directly from examples, data, and experience" (The Royal Society 2017, 16); deep learning, considered a subfield of machine learning, is roughly patterned on the neural networks present in the human brain — such that there are networks of artificial neurons used in the processing of data. The second driver of this change is the vast quantity of data that can be employed for training, combined with an exponential increase in computing power.

In their current, and likely future, iterations, these technologies present policy makers with a number of dilemmas. When technology can learn for itself, "think" for itself and — when combined with autonomous robotics — ultimately do for itself, the governance challenges become complex. There are ethical questions, and problems around explainability, safety, reliability and accountability, to name a few.

In the series of essays that follows, international experts seek to make assessments of the near-, medium- and long-term policy implications of the increased deployment of AI and autonomous systems in military and security applications, recognizing (of course) that the further the time horizon is extended, the more abstract and speculative the analysis becomes. The series also seeks to address some of the big, looming policy questions:

- Is existing international law adequate?
- Will this technology upend or change traditional state power structures?
- Will AI be a stabilizing or a destabilizing force in international relations?
- How will states work with the private sector, and vice versa, and what impact will those decisions have?

As this series of essays makes clear, the most significant and complicated governance challenges related to the deployment of AI are in the areas of defence, national security and international relations. In this context, Canada's defence policy, *Strong, Secure, Engaged*, laid the problem bare: "State and non-state actors are increasingly pursuing their agendas using hybrid methods in the 'grey zone' that exists just below the threshold of armed conflict. Hybrid methods involve the coordinated application of diplomatic, informational, cyber, military and economic instruments to achieve strategic or operational objectives. They often rely on the deliberate spread of misinformation to sow confusion and discord in the international community, create ambiguity and maintain deniability. The use of hybrid methods increases the potential for misperception and miscalculation" (National Defence Canada 2017, 53).

This suite of challenges is set to be magnified as AI gets better and better, and as adversarial actors continue to lurk below that threshold of traditional armed conflict. Inevitably, these challenges will continue to put pressure on the existing international rules and governance structures, many of which were designed for a different era. The moment is right to contemplate innovative policy solutions to match these challenges. We hope that the thinking advanced in this essay series may offer some guidance for policy makers in these extremely challenging times.

Promoting Innovation While Offering Practical and Flexible Regulation

As Daniel Araya and Rodrigo Nieto-Gómez tell us in their essay "Renewing Multilateral Governance in the Age of AI," the most challenging part of developing AI policy and regulatory regimes is identifying what, specifically, must be regulated. This, they note, is due to the fact that AI technologies are designed not as end products, but rather as "ingredients or components within a wide range of products, services and systems," ultimately encouraging the proliferation of combinatorial technologies. Therefore, as Araya and Nieto-Gómez suggest, successful regulation is "less about erecting non-proliferation regimes…and more about creating good design norms and principles" that carefully weigh design concerns against technical and ethical ones.

So, where should policy makers begin? In her essay "Public and Private Dimensions of AI Technology and Security," Maya Medeiros suggests that individual governments will necessarily take the lead on a national level, but must coordinate "effort between different public actors. Different regulators with similar policy objectives should adopt universal language for legislation to encourage regulatory convergence."

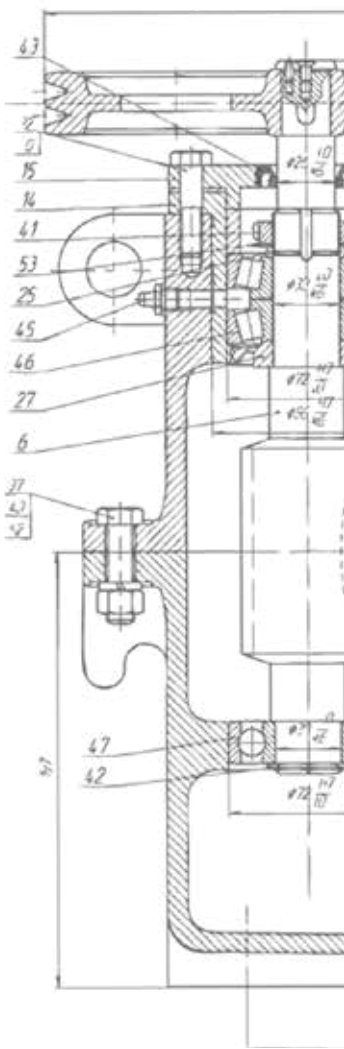
There is good news: We do not have to reinvent a regulatory regime for AI. Many of the sectors that will be significantly impacted by AI already have a strong regulatory history. Consider the automotive sector. In the United States, the National Highway Traffic Safety Administration regulates the safety of automobiles, and the Environmental Protection Agency regulates vehicle emissions, while state and local governments are able to establish their own safety laws and regulations as long as they do not conflict with federal standards.

But, even as specific sectors take different approaches to regulating AI capabilities, new AI norms, laws and regulations need to be general enough so that they do not become outdated. As Liis Vihul urges in her essay "International Legal Regulation of Autonomous Technologies," they cannot be so vague that they are useless. For the defence sector, there is already a place to start: Vihul describes a "dense international legal framework" for warfare in which many rules may already exist that regulate the use of autonomous technologies in conflict. However, we will learn only in the application of these tools whether the old framework holds up.

Security Challenges Ahead

As if the challenge of governing AI was not hard enough, Michael C. Horowitz suggests, in "AI and the Diffusion of Global Power," that AI will necessarily make cybersecurity threats a lot more complicated. He argues that cyberespionage could evolve to focus on algorithm theft, while data poisoning could prevent adversaries from developing effective algorithms. Amandeep Singh Gill, in his essay "A New Arms Race and Global Stability," specifically identifies the spoofing of image recognition models by adversarial attackers as one such risk, which would only make the disinformation challenges outlined by Samantha Bradshaw that much more complicated for governments to address. These challenges are, collectively, as Bradshaw points out in her essay "Influence Operations and Disinformation on Social Media," a systems problem. So, rather than simply labelling the content as a problem, we need to find a way toward a solution — starting with acknowledging that social media platforms have both the responsibility and the technical agency to effectively moderate our information ecosystems.

Finally, the risk of miscalculation looms as possibly the most significant threat posed by AI, because these tools are so new that we have yet to develop robust policy, legal and process frameworks for protecting against their misuse, intentional or unintentional. In his essay "Artificial Intelligence and Keeping Humans 'in the Loop,'" Robert Mazzolin assesses whether, when and how humans should be kept in the decision-making loop; he suggests that decisions will depend on how adept AI becomes at the crucial tasks of discriminating between different data sets to properly "self-learn," and noticing attempts at manipulation. Similarly, Gill argues that we need to advocate for the "gift of time" to safeguard the use of autonomous weapons.



Next Steps for Policy Makers

It would be easy to feel overwhelmed by the task of governing emerging AI technologies. Thankfully, this series' contributing authors have laid out a blueprint for assessing catastrophic risk, supporting peaceful and commercial use, and preserving North American technological leadership for near-term AI applications. Four key recommendations emerge that policy makers can apply in addressing the expanded use of AI and its highly unpredictable potential.

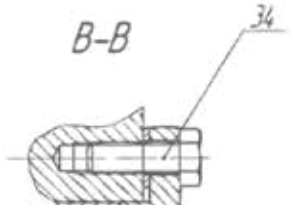
First, policy makers must prioritize developing a multidisciplinary network of trusted experts on whom to call regularly to identify and discuss new developments in AI technologies, many of which may not be intuitive or even yet imagined. In the same way that Marvin Minsky could not have predicted 70 years ago what AI would be capable of today, the one certainty for today's policy makers is uncertainty. The large-scale deployment of AI, especially in security contexts, presents a range of horizontal public policy issues — international trade, intellectual property, data governance, domestic innovation strategy and national security, to name a few. Given the often complex interrelationships within and among these areas of concern, having access to multidisciplinary expertise will be a must.

Second, policy makers must work to develop strategies that are flexible enough to accommodate tactical shifts when the technology advances — for example, as computing power and algorithmic quality improve — and that allow for system-level changes. The policy frameworks they develop must be capable of attenuating the potential negative aspects of AI, while also maintaining enough elasticity to account for the inevitable advancement in technology capability and expanded use cases.

Third, policy makers must invest significant time and resources — in close cooperation with the private sector — to identifying the specific AI applications most in need of governance frameworks. This work must principally include continuously assessing each of AI's many subsectors to identify the relative technological advancements of specific nations and regions. It will also require that policy makers move quickly (but deliberately) in certain areas where the storm is already upon us — one need only consider the interplay between behavioural nudging, big data, personal data, micro-targeting, feedback loops and foreign adversarial influence on everything from elections to societal cohesion.

Fourth, working in tandem with existing international regulatory bodies, policy makers must ensure not only that universal AI governance frameworks are consistent with their respective national regulations, but also that foundational principles — notably, human rights — are respected at every stage from design to implementation. To put it bluntly, these principles, which include rights to privacy, freedom of thought and conscience, among others, are too important to trade away for the sake of design. Safeguarding them will require policy makers to remain vigilant and to better understand the geostrategic elements of technical design, international standard setting and market development, because these are areas where adversarial states are always seeking advantage.

While narrow AI systems will likely continue to outperform their human counterparts, there is little evidence to suggest that these applications, as sophisticated as they may be, will evolve rapidly into systems with general intelligence.



Conclusion

As challenging as this moment may be, it offers a significant opportunity for policy makers; it is critical to remember, as Vihul points out, that "autonomous technologies are in their infancy." Today and in the near future, we are talking only about "narrow" AI applications such as those derived through statistical machine learning. If, as Horowitz argues, artificial general intelligence — or machines capable of carrying out a broad range of cognitive tasks as well as or better than humans — is achieved, the governance playbook will need to be revised.

While narrow AI systems will likely continue to outperform their human counterparts, there is little evidence to suggest that these applications, as sophisticated as they may be, will evolve rapidly into systems with general intelligence. For example, a recent test by *MIT Technology Review* of Open AI's new language-generating model, GPT-3, displayed the model's "poor grasp of reality" despite its impressive "175 billion parameters and 450 gigabytes of input data," with the reviewers concluding it "does not make for trustworthy intelligence" (Marcus and Davis 2020).

Put simply, AI applications have, and will continue to have, significant limitations, and those limitations must be accounted for as systems of governance are designed.

Instead, disruption will more likely happen in the combination of technologies — robotics and AI, for example. Identifying these trend lines, and being able to offer flexible but specific-enough policy within adaptable regulatory and legal frameworks — essentially, governance guardrails — that can respond when technology does evolve, will be critical for ensuring the new dimensions of international security.

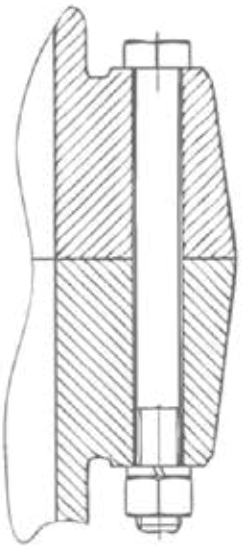
As adversarial states continue to engage in the use of hybrid methods in the "grey zone," policy makers can expect the challenges to become more pronounced as AI technology continues its rapid development. They can also expect that modern conflict and the future battlespace will be profoundly entangled with AI and autonomous systems. As the world moves into a deeply fragmented time, defined by distrust and great power competition, AI holds the potential to be a destabilizing force that can increase the likelihood of miscalculation, if it is deployed without adequate governance mechanisms in place.

WORKS CITED

Marcus, Gary and Ernest Davis. 2020. "GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about." *MIT Technology Review*, August 22. www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/.

National Defence Canada. 2017. *Strong, Secure, Engaged: Canada's Defence Policy*. Ottawa, ON: National Defence. http://dgpaapp.forces.gc.ca/en/canada-defence-policy/docs/canada-defence-policy-report.pdf.

The Royal Society. 2017. *Machine learning: the power and promise of computers that learn by example*. Full report (DES4702), April. London, UK: The Royal Society. https://royalsociety.org/machine-learning.



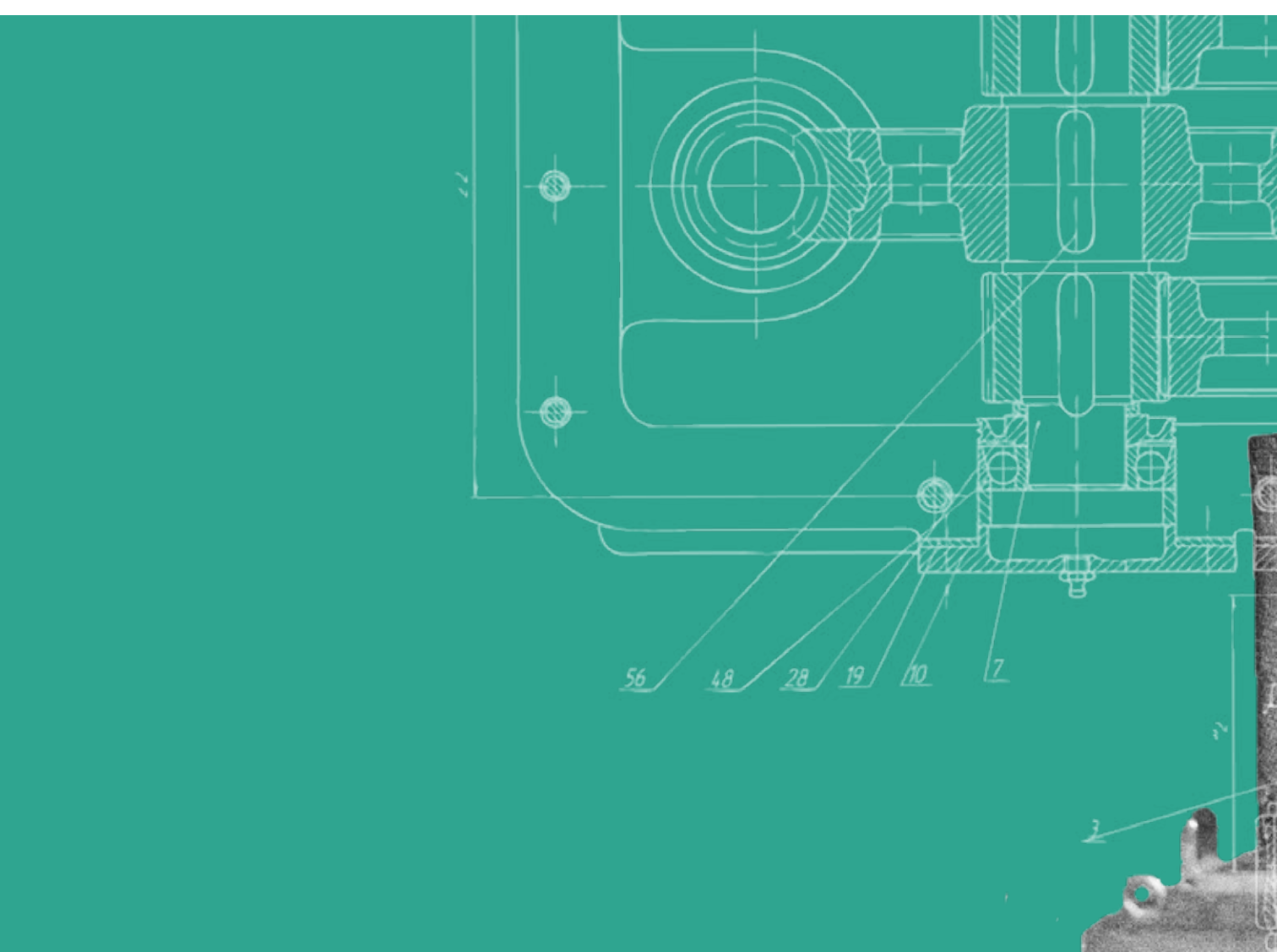
ABOUT THE AUTHORS

Meg King is the director of the Wilson Center's Science and Technology Innovation Program and leads cutting-edge training programs to equip generations of US Congressional and Executive branch staff with technology skills. Previously, she was an international project manager for the US government's Cooperative Threat Reduction Program. From 2008 to 2011, she was senior legislative assistant to the chair of the House Homeland Security Subcommittee on Intelligence, Information Sharing, and Terrorism Risk Assessment. She began her career as a research assistant for the Homeland Security Project at the Center for the Study of the Presidency. Meg is a member of the International Institute for Strategic Studies and of the Women in International Security Network.

As CIGI's managing director and general counsel, Aaron Shull acts as a strategic liaison between CIGI's research initiatives and other departments while managing CIGI's legal affairs and advising senior management on a range of legal, operational and policy matters. A member of CIGI's executive team, Aaron provides guidance and advice on matters of strategic and operational importance, while working closely with partners and other institutions to further CIGI's mission. He also serves as corporate secretary. Aaron is an expert on cybersecurity issues, and coordinated the CIGI essay series Governing Cyberspace during a Crisis in Trust. Prior to joining CIGI, Aaron practised law for a number of organizations, focusing on international, regulatory and environmental law.

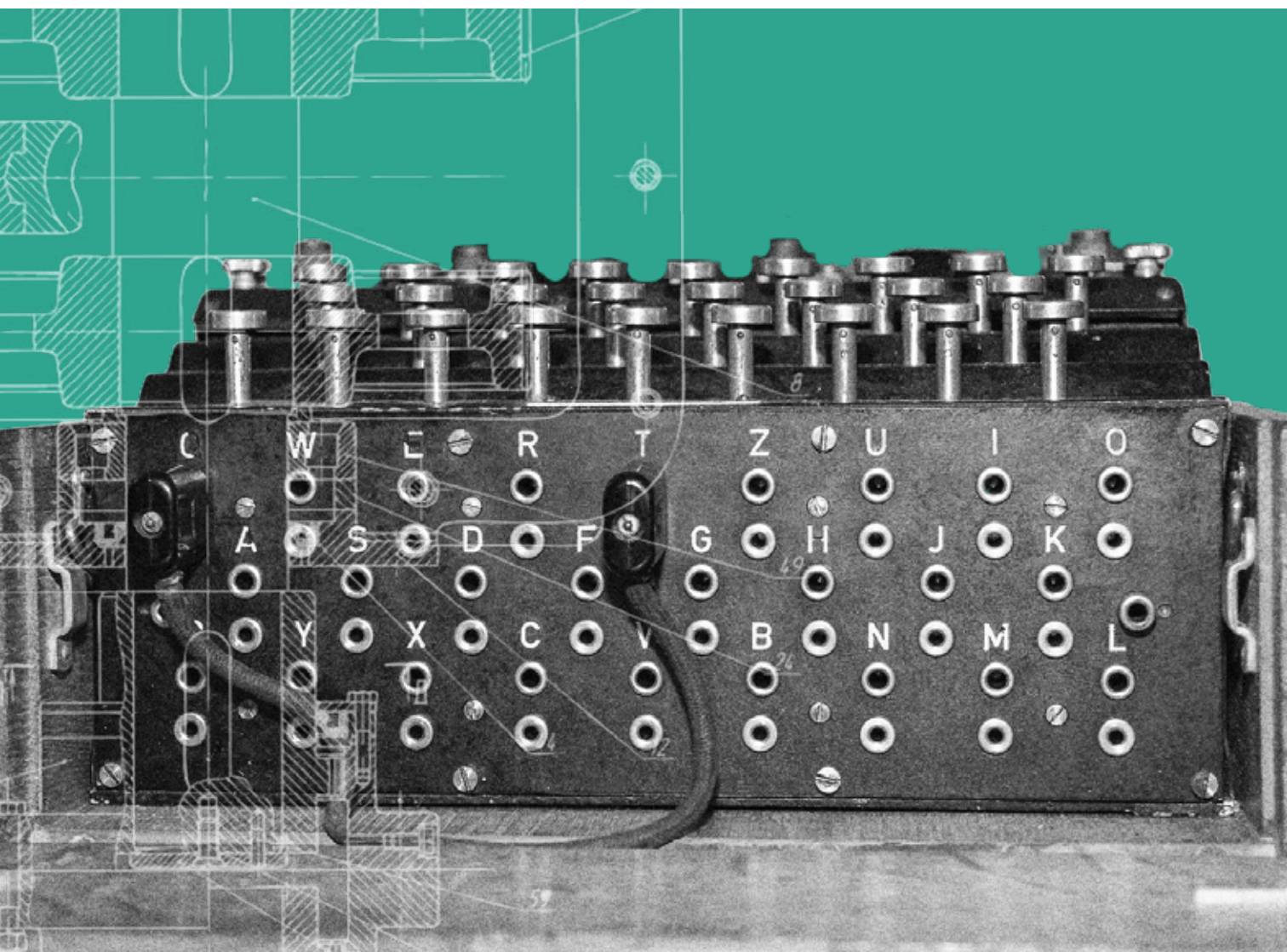
Renewing Multilateral Governance in the Age of AI

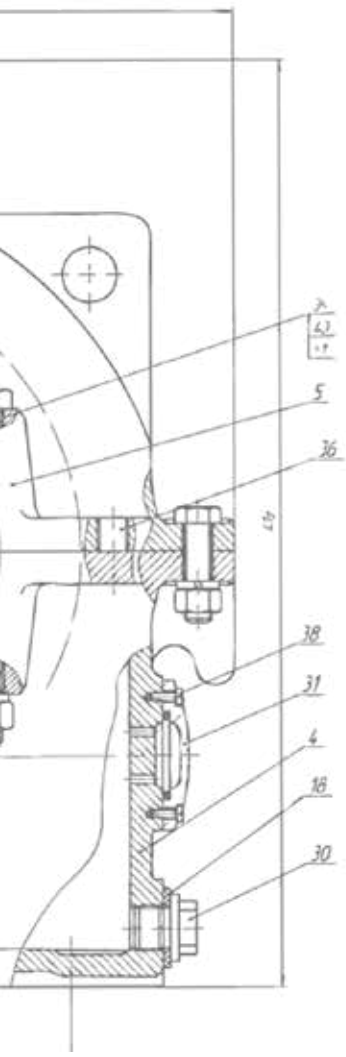
Daniel Araya and Rodrigo Nieto-Gómez



Since its inception some 60 years ago (Anyoha 2017), artificial intelligence (AI) has evolved from an arcane academic field into a powerful driver of social transformation. AI and machine learning are now the basis for a wide range of mainstream commercial applications, including Web search (Metz 2016), medical diagnosis (Davis 2019), algorithmic trading (*The Economist* 2019), factory automation (Stoller 2019), ridesharing (Koetsier 2018) and, most recently, autonomous vehicles (Elezaj 2019). Deep learning — a form of machine learning — has dramatically improved pattern recognition, speech recognition and natural language processing (NLP). But AI is also the basis for a highly competitive geopolitical contest.

Much as mass electrification accelerated the rise of the United States[1] and other advanced economies, so AI has begun reshaping the contours of the global order. Data-driven technologies are now the core infrastructure around which the global economy operates. Indeed, whereas intangible assets (patents, trademarks and copyrights) were only 16 percent of the S&P 500 in 1976, they comprise 90 percent today (Ocean Tomo 2020, 2). In fact, intangible assets for S&P 500 companies are worth a staggering US$21 trillion (Ross 2020; see also Table 1). In this new era, power is rooted in technological innovation. Together, renewable energy technologies





(Araya 2019b), fifth-generation (5G) telecommunications (Araya 2019a), the Industrial Internet of Things (Wired Insider 2018) and, most importantly, AI are now the foundations for a new global order.

At the research level, the United States remains the world's leader in AI. The National Science Foundation (NSF) provides more than US$100 million each year in AI funding (NSF 2018). The Defense Advanced Research Projects Agency (DARPA) has recently promised US$2 billion in investments toward next-generation AI technologies (DARPA 2018). In fact, the United States leads in the development of semiconductors, and its universities train the world's best AI talent. But, while the United States retains a significant research dominance, national leadership is weak and resources are uneven.

While the United States has established a strong lead in AI discovery, it is increasingly likely that China could dominate AI's commercial application (Lee 2018). China has now emerged as an AI powerhouse (Simonite 2019), with advanced commercial capabilities in AI and machine learning (Lee 2020) and a coherent national strategy. Alongside China's expanding expertise in factory machinery, electronics, infrastructure and renewable energy, Beijing has made AI a top priority (Allen 2019). What is obvious is that China has begun a long-term strategic shift around AI and other advanced technologies (McBride and Chatzky 2019).

Toward a New Global Order

Together, China and the United States are emerging as geopolitical anchors for a new global order (Anderson 2020). But what kind of global order? How should actors within the existing multilateral system anticipate risks going forward? Even as the current coronavirus disease 2019 (COVID-19) crisis engulfs a wide range of industries and institutions, technological change is now set to undermine what is left of America's Bretton Woods system (Wallace 2020). To be sure, weaknesses in the current multilateral order threaten to deepen the world's geopolitical tensions.

The truth is that a global system shift is already under way (Barfield 2018). As Rohinton P. Medhora and Taylor Owen (2020) have pointed out, data-driven technologies and the current pandemic are both manifestations of an increasingly unstable system. For almost five decades, the United States has guided the growth of an innovation-driven order (Raustiala 2017). But that order is coming to an end. Accelerated by the current health crisis, the world economy is now fragmenting (Pethokoukis 2020). Beyond the era of US hegemony, what we are increasingly seeing is a rising techno-nationalism that strategically leverages the network effects of technology to reshape a post-Bretton Woods order (Rajan 2018).

Table 1: Largest Companies by Intangible Value

Rank	Company	Sector	Total Intangible Value (in billion US$)	Share of Enterprise Value
1	Microsoft	Internet and software	$904	90%
2	Amazon	Internet and software	$839	93%
3	Apple	Technology and information technology	$675	77%
4	Alphabet	Internet and software	$521	65%
5	Facebook	Internet and software	$409	79%
6	AT&T	Telecommunications	$371	84%
7	Tencent	Internet and software	$365	88%
8	Johnson & Johnson	Pharma	$361	101%
9	Visa	Banking	$348	100%
10	Alibaba	Internet and software	$344	86%

Source: Ross (2020), with data from Brand Finance plc (2019, 18).
Note: Percentages may exceed 100% due to rounding.

One could observe (as Vladimir Lenin purportedly did), that "there are decades where nothing happens, and there are weeks where decades happen." Even as nation-states leverage data to compete for military and commercial advantage, disruptive technologies such as AI and machine learning are set to transform the nature and distribution of power. What seems clear is that we are entering an era of hybrid opportunities and challenges generated by the combination of AI and a cascading economic crisis. The spread of smart technologies across a range of industries suggests the need for rethinking the institutions that now govern us (Boughton 2019).

In order to manage rising tensions, a new and coordinated global governance framework for overseeing AI is needed. The absence of effective global governance in this new era means that we are facing significant turbulence ahead. Indeed, in the wake of COVID-19, the global economy could contract by as much as eight percent, moving millions of people into extreme poverty (World Bank 2020). Any new framework for multilateral governance will need to oversee a host of challenges overlapping trade, supply-chain reshoring, cyberwar, corporate monopoly, national sovereignty, economic stratification, data governance, personal privacy and so forth.

Global Governance in the AI Era

Perhaps the most challenging aspect of developing policy and regulatory regimes for AI is the difficulty in pinpointing what exactly these regimes are supposed to regulate. Unlike nuclear proliferation or genetically modified pathogens, AI is not a specific technology; it's more akin to a collection of materials than a particular weapon. It is also an aspirational goal (Walch 2018), much like a philosopher's stone that drives the magnum opus of computer science (the agent that unlocks the alchemy). To take only one example, Peter J. Denning and Ted G. Lewis (2019) classify the idea of "sentient" AI as largely "aspirational."

The dream of the intelligent machine now propels computer science, and therefore regulatory systems, around the world. Together, research in AI and aspirational expectations around sentient machines are now driving fields as diverse as image analysis, automation, robotics, cognitive and behavioural sciences, operations research and decision making, language processing and the video gaming industry (The Verge 2019) — among many others. Regulating AI, therefore, is less about erecting non-proliferation regimes (a metaphor often used for managing AI; see, for example, Frantz 2018), and more about creating good design norms and principles that "balance design trade-offs not only among technical constraints but also among ethical constraints" (Ermer and VanderLeest 2002, 7.1253.1), across all sorts of products, services and infrastructure.

The spread of smart technologies across a range of industries suggests the need for rethinking the institutions that now govern us.

To explain this challenge with greater precision, we need to better appreciate the truism that "technology is often stuff that doesn't work yet" and apply a version of this truism to the discourse on AI as "the stuff computers still can't do." In 1961, the development of a computer "spelling checker" (today a ubiquitous application) fell to the inventors at the Stanford Artificial Intelligence Laboratory (Earnest n.d.), like AI. Today, it is "just" a software feature. Twenty years ago, a car accelerating, braking and parking itself would have been considered AI; today, those functions are "just" assisted driving. Google is a search engine and Uber a ride-sharing app, but few of the billions of users of these Web platforms would consider them AI.

Simply put, once a service or a product successfully integrates AI research into its value proposition, that technology becomes a part of the functionality of a system or service. Meanwhile, the application of this "weak" or "narrow" AI proliferates across multiple research disciplines in the diffusion of an expanding horizon of tools and technologies. These applications of AI are everywhere and, in becoming everyday "stuff that works," simply "disappear" into the furniture and functionality

of the systems and services they augment. To take one example, a chair with adaptive ergonomics based on machine learning is just a fancy chair, not AI. For this reason, it would be very difficult to build general-purpose regulatory regimes that anticipate every case of AI and machine learning. The innovation costs alone of trying to do so would be staggering.

Toward Mundane AI Regulation

The challenges in regulating AI are, then, twofold: On the one hand, if we understand AI as a series of technological practices that replicate human activities, then there is simply no single field to regulate. Instead, AI governance overlaps almost every kind of product or service that uses computation to perform a task. On the other hand, if we understand AI and dedicated AI laboratories as the basis for dramatically altering the balance of power among peoples and nations, then we have terrific challenges ahead. Beyond the exaggerations of AI often seen in science fiction, it is clearly important to develop the appropriate checks and balances to limit the concentration of power that AI technologies can generate.

The need for AI regulation has opened a Pandora's box of challenges that cannot be closed.

Instead of the mostly binary nuclear non-proliferation lens often used to discuss AI governance, inspiration for a more relevant (albeit less exciting) model of regulation can be found in food regulation (specifically, food safety) and material standards. Much like the products and processes falling within these two regulatory domains, AI technologies are designed not as final entities, but as ingredients or components to be used within a wide range of products, services and systems. AI algorithms, for example, serve as "ingredients" in the combinatorial technologies. These technologies include search engines (algorithmic ranking), military drones (robotics and decision making) and cybersecurity software (algorithmic optimization). But they also include mundane industries such as children's toys (for semantic analysis, visual analysis and robotics[2]) and social media networks (for trend analysis and predictive analytics; see, for example, Rangaiah 2020).

The need for AI regulation has opened a Pandora's box of challenges that cannot be closed. And should not be — just as the quest for the unattainable philosopher's stone created some of the most important foundational knowledge in modern chemistry (Hudson 1992), so the search for "strong AI" has produced some of the core ingredients used to develop the most exciting, profitable and powerful modern technologies that now exist. In this sense, AI technologies behave less like nuclear technologies and more like aspartame or polyethylene.

Fortunately, the mature regulatory regimes overseeing food safety and material standards have already produced a series of norms that can inspire the ways in which global AI regulation might work. Instead of trying to regulate the function or final shape of an AI-enabled technology, the object of regulation should instead focus on AI as an "ingredient" or component of technological proliferation. This approach will be particularly important in preserving innovation capacity while providing appropriate checks and balances on the proliferation of AI-driven technologies.

Envisioning Smart AI Governance

Notwithstanding the mundane aspects of AI governance, very real challenges lie ahead. We are living through a period of transition between two epochs: an industrial era characterized by predictable factory labour, and a new digital era characterized by widespread institutional unravelling. In this new century, the United States remains a formidable power, but its days of unipolar hegemony have come to an end. The hard reality is that technology is disrupting the geopolitical and regulatory landscape, driving the need for new protocols and new regulatory regimes.

Without a doubt, the most complex governance challenges surrounding AI today involve defence and security. From killer swarms of drones (Future of Life Institute 2017) to the computer-assisted enhancement of the military decision-making process

(Branch 2018), AI technologies will force multiply the capacity of nation-states to project power. While the temptation to use the non-proliferation lens for any other kind of AI technology (for example, ban all killer robots!), the dual-use challenge remains the same. A killer robot is not a specific kind of technology. It is, instead, the result of the recombination of AI "ingredients," many of which are also used to, for example, detect cancers or increase driver safety.

Over and above the current COVID-19 crisis, data-driven technologies are provoking a vast geotechnological restructuring (Khanna 2014). In this new environment, AI and machine learning are set to reshape the rules of the game. As Google's Sundar Pichai (2020) wrote early this year in an op-ed for the *Financial Times*, the time for properly regulating AI technologies is now. As in the postwar era, what we need is a new kind of multilateral system to oversee a highly technological civilization. Sadly, much of the existing governance architecture lacks the capacity to address the needs of a data-driven economy. Nonetheless, most governments are already beginning to explore new regulations, even as approaches vary.

Given the scale of the changes ahead, we will need to consider the appropriate regimes for regulating AI. Fortunately, this does not mean starting from scratch. Even as regulatory compliance issues around AI proliferate, many existing regulatory systems and frameworks will remain invaluable. Indeed, even as the final forms of many AI technologies differ, the underlying ingredients are shared. And just as consumer protection laws hold manufacturers, suppliers and retailers accountable, so the plethora of AI-driven products and services can be similarly regulated. Nonetheless, looking beyond the mundane regulation of AI, many big challenges remain. Solving these challenges will mean rethinking a waning multilateral order.

NOTES

1 See www.instituteforenergyresearch.org/history-electricity/.

2 See, for example, https://embodied.com/products/moxie-reservation.

WORKS CITED

Allen, Gregory C. 2019. "Understanding China's AI Strategy: Clues to Chinese Strategic Thinking on Artificial Intelligence and National Security." Center for a New American Security, February 6. www.cnas.org/publications/reports/understanding-chinas-ai-strategy.

Anderson, Bruce. 2020. "The new world order is disorder." *Canada's National Observer*, May 25. www.nationalobserver.com/2020/05/25/opinion/new-world-order-disorder.

Anyoha, Rockwell. 2017. "Can Machines Think?" *Science in the News* (blog), August 28. Boston, MA: Harvard University Graduate School of Arts and Sciences. http://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/.

Araya, Daniel. 2019a. "Huawei's 5G Dominance In The Post-American World." *Forbes*, April 5. www.forbes.com/sites/danielaraya/2019/04/05/huaweis-5g-dominance-in-the-post-american-world/.

———. 2019b. "The Green New Deal: Jeremy Rifkin And The Coming Collapse." *Forbes*, October 8. www.forbes.com/sites/danielaraya/2019/10/08/the-green-new-deal-jeremy-rifkin-and-the-coming-collapse/.

Barfield, Claude. 2018. "Will the US lose by default in competition for digital trade rules?" *AEIdeas* (blog), August 15. Washington, DC: American Enterprise Institute. www.aei.org/technology-and-innovation/digital-trade/will-the-us-lose-by-default-in-competition-for-digital-trade-rules/.

Boughton, James. 2019. "Can We Expect a 'New Bretton Woods'?" The Bretton Woods Committee, February 5. www.brettonwoods.org/article/can-we-expect-a-new-bretton-woods.

Branch, William A. 2018. "Artificial Intelligence and Operational-Level Planning: An Emergent Convergence." Master's thesis, US Army Command and General Staff College. https://apps.dtic.mil/dtic/tr/fulltext/u2/1070958.pdf.

Brand Finance plc. 2019. *Global Intangible Finance Tracker 2019 — an annual review of the world's intangible value.* November. London, UK: Brand Finance. https://brandfinance.com/images/upload/gift_2.pdf.

DARPA. 2018. "DARPA Announces $2 Billion Campaign to Develop Next Wave of AI Technologies." DARPA News and Events, September 7. www.darpa.mil/news-events/2018-09-07.

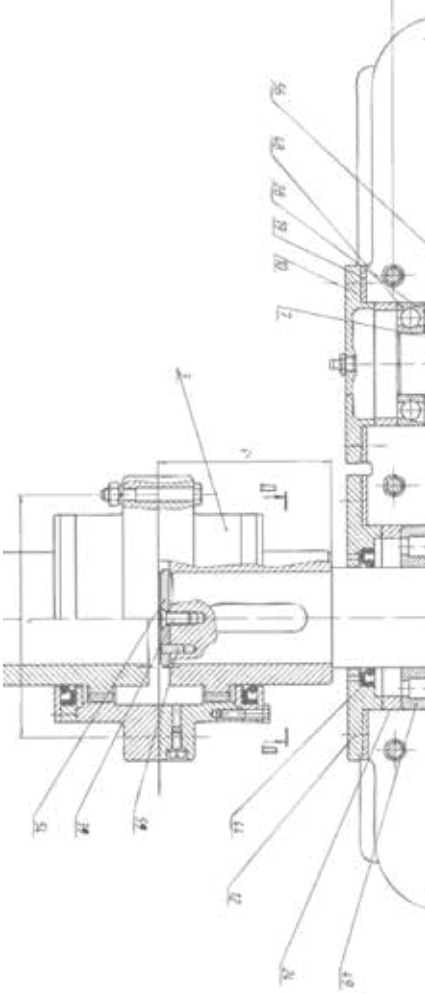
Davis, Nicola. 2019. "AI equal with human experts in medical diagnosis, study finds." *The Guardian*, September 24. www.theguardian.com/technology/2019/sep/24/ai-equal-with-human-experts-in-medical-diagnosis-study-finds.

Denning, Peter J. and Ted G. Lewis. 2019. "Intelligence May Not Be Computable." *American Scientist* 107 (6): 346. www.americanscientist.org/article/intelligence-may-not-be-computable.

Earnest, Les. n.d. "SAIL Away." Reprint of "HELLO SAILOR!" published 1995 in *The Analytical Engine.* https://web.stanford.edu/~learnest/sail/sailaway.htm.

Elezaj, Rilind. 2019. "How AI Is Paving the Way for Autonomous Cars." *Machine Design*, October 17. www.machinedesign.com/mechanical-motion-systems/article/21838234/how-ai-is-paving-the-way-for-autonomous-cars.

Ermer, Gayle E. and Steven H. VanderLeest. 2002. "Using Design Norms to Teach Engineering Ethics." Paper presented at the 2002 American Society for Engineering Education Annual Conference & Exposition, Montreal, Canada, June 16–19. https://peer.asee.org/using-design-norms-to-teach-engineering-ethics.



Frantz, Douglas. 2018. "Op-Ed: We've unleashed AI. Now we need a treaty to control it." *Los Angeles Times*, July 16. www.latimes.com/opinion/op-ed/la-oe-frantz-artificial-intelligence-treaty-20180716-story.html.

Future of Life Institute. 2017. "Slaughterbots." November 17. YouTube video, 7:47. www.youtube.com/watch?v=HipTO_7mUOw.

Hudson, John. 1992. "From Alchemy to Chemistry." In *The History of Chemistry*, 35–46. Boston, MA: Springer. https://doi.org/10.1007/978-1-4684-6441-2_3.

Khanna, Parag. 2014. "Geotechnology and Global Change." *Global Policy* 5 (1): 56. https://doi.org/10.1111/1758-5899.12117.

Koetsier, John. 2018. "Uber Might Be the First AI-First Company, Which Is Why They 'Don't Even Think About It Anymore.'" *Forbes*, August 22. www.forbes.com/sites/johnkoetsier/2018/08/22/uber-might-be-the-first-ai-first-company-which-is-why-they-dont-even-think-about-it-anymore/#5f88a1435b62.

Lee, Kai-Fu. 2018. "Why China Can Do AI More Quickly and Effectively Than the US." *Wired*, October 23. www.wired.com/story/why-china-can-do-ai-more-quickly-and-effectively-than-the-us/.

———. 2020. "Kai-Fu Lee on how covid spurs China's great robotic leap forward." *The Economist*, June 25. www.economist.com/by-invitation/2020/06/25/kai-fu-lee-on-how-covid-spurs-chinas-great-robotic-leap-forward.

McBride, James and Andrew Chatzky. 2019. "Is 'Made in China 2025' a Threat to Global Trade?" Backgrounder, May 13. New York, NY: Council on Foreign Relations. www.cfr.org/backgrounder/made-china-2025-threat-global-trade.

Medhora, Rohinton P. and Taylor Owen. 2020. "A Post-COVID-19 Digital Breton Woods." Project Syndicate, April 17; CIGIonline, April 19. www.cigionline.org/articles/post-covid-19-digital-bretton-woods.

Metz, Cade. 2016. "AI Is Transforming Google Search. The Rest of the Web Is Next." *Wired*, February 4. www.wired.com/2016/02/ai-is-changing-the-technology-behind-google-searches/.

NSF. 2018. "Statement on Artificial Intelligence for American Industry." NSF Press Statement 18-005, May 10. www.nsf.gov/news/news_summ.jsp?cntn_id=245418.

Ocean Tomo. 2020. "Intangible Asset Market Value Study." Interim study update, July 1. www.oceantomo.com/intangible-asset-market-value-study/.

Pethokoukis, James. 2020. "Artificial intelligence and the post-pandemic economy: My long-read Q&A with Roger Bootle." *AEIdeas* (blog), April 23. Washington, DC: American Enterprise Institute. www.aei.org/economics/artificial-intelligence-and-the-post-pandemic-economy-my-long-read-qa-with-roger-bootle/.

Pichai, Sundar. 2020. "Why Google thinks we need to regulate AI." *Financial Times*, January 19. www.ft.com/content/3467659a-386d-11ea-ac3c-f68c10993b04.

Rajan, Amol. 2018. "Techno-nationalism could determine the 21st Century." BBC News, September 8. www.bbc.com/news/technology-45370052.

Rangaiah, Mallika. 2020. "How is Artificial Intelligence (AI) Making TikTok Tick?" *Analytics Steps* (blog), January 16. www.analyticssteps.com/blogs/how-artificial-intelligence-ai-making-tiktok-tick.

Raustiala, Kal. 2017. "An Internet Whole and Free: Why Washington Was Right to Give Up Control." *Foreign Affairs*, March/April. www.foreignaffairs.com/articles/world/2017-02-13/internet-whole-and-free.

Ross, Jenna. 2020. "Intangible Assets: A Hidden but Crucial Driver of Company Value." *Visual Capitalist*, February 11. www.visualcapitalist.com/intangible-assets-driver-company-value/.

Simonite, Tom. 2019. "China Is Catching Up to the US in AI Research — Fast." *Wired*, March 13. www.wired.com/story/china-catching-up-us-in-ai-research/.

Stoller, Jacob. 2019. "Artificial intelligence meets the real world." *Manufacturing Automation*, March 18. www.automationmag.com/artificial-intelligence-meets-the-real-world-9187/.

The Economist. 2019. "The stockmarket is now run by computers, algorithms and passive managers." Briefing, October 5. www.economist.com/briefing/2019/10/05/the-stockmarket-is-now-run-by-computers-algorithms-and-passive-managers.

The Verge. 2019. "How AI will completely change video games." March 6. YouTube video, 7:08. www.youtube.com/watch?v=NPuYtHZud0o&feature=youtu.be.

Walch, Kathleen. 2018. "Artificial Intelligence Is Not A Technology." *Forbes*, November 1. www.forbes.com/sites/cognitiveworld/2018/11/01/artificial-intelligence-is-not-a-technology/?sh=173ae6975dcb.

Wallace, Tim. 2020. "Why the world needs a new Bretton Woods moment." *The Telegraph*, May 27. www.telegraph.co.uk/business/2020/05/27/world-needs-new-bretton-woods-moment/.

Wired Insider. 2018. "Industrial IoT: How Connected Things Are Changing Manufacturing." www.wired.com/wiredinsider/2018/07/industrial-iot-how-connected-things-are-changing-manufacturing/.

World Bank. 2020. "COVID-19 to Plunge Global Economy into Worst Recession since World War II." Press release, June 8. www.worldbank.org/en/news/press-release/2020/06/08/covid-19-to-plunge-global-economy-into-worst-recession-since-world-war-ii.

ABOUT THE AUTHORS

Daniel Araya is a CIGI senior fellow, a senior partner with the World Legal Summit, and a consultant and an adviser with a special interest in artificial intelligence, technology policy and governance. At CIGI, his work contributes to research on autonomous systems in global governance and looks specifically at the best ways to mitigate the negative effects of the widespread deployment of new technologies. Daniel is a regular contributor to various media outlets and organizations such as Forbes, the Brookings Institution, Futurism and Singularity Hub. He has been invited to speak at a number of universities and research centres, including the US Naval Postgraduate School; Harvard University; the American Enterprise Institute; the Center for Global Policy Solutions; Stanford University; the University of Toronto; the University of California, Santa Cruz; and Microsoft Research. His most recent books include *Augmented Intelligence: Smart Systems and the Future of Work and Learning* (2018) and *Smart Cities as Democratic Ecologies* (2015). Daniel has a doctorate from the University of Illinois at Urbana-Champaign.

Rodrigo Nieto-Gómez is a geostrategist and defence futurist focused on the consequences of the accelerating pace of change in homeland security and policing environments. He is a research professor at the National Security Affairs Department and at the Center for Homeland Defense and Security, both at the Naval Postgraduate School in Monterey, California. He has also worked as a certified facilitator and instructor for the Command College for the California Commission on Peace Officer Standards and Training and instructed at the Executive Academy of the Emergency Management Institute.

big · tech

with

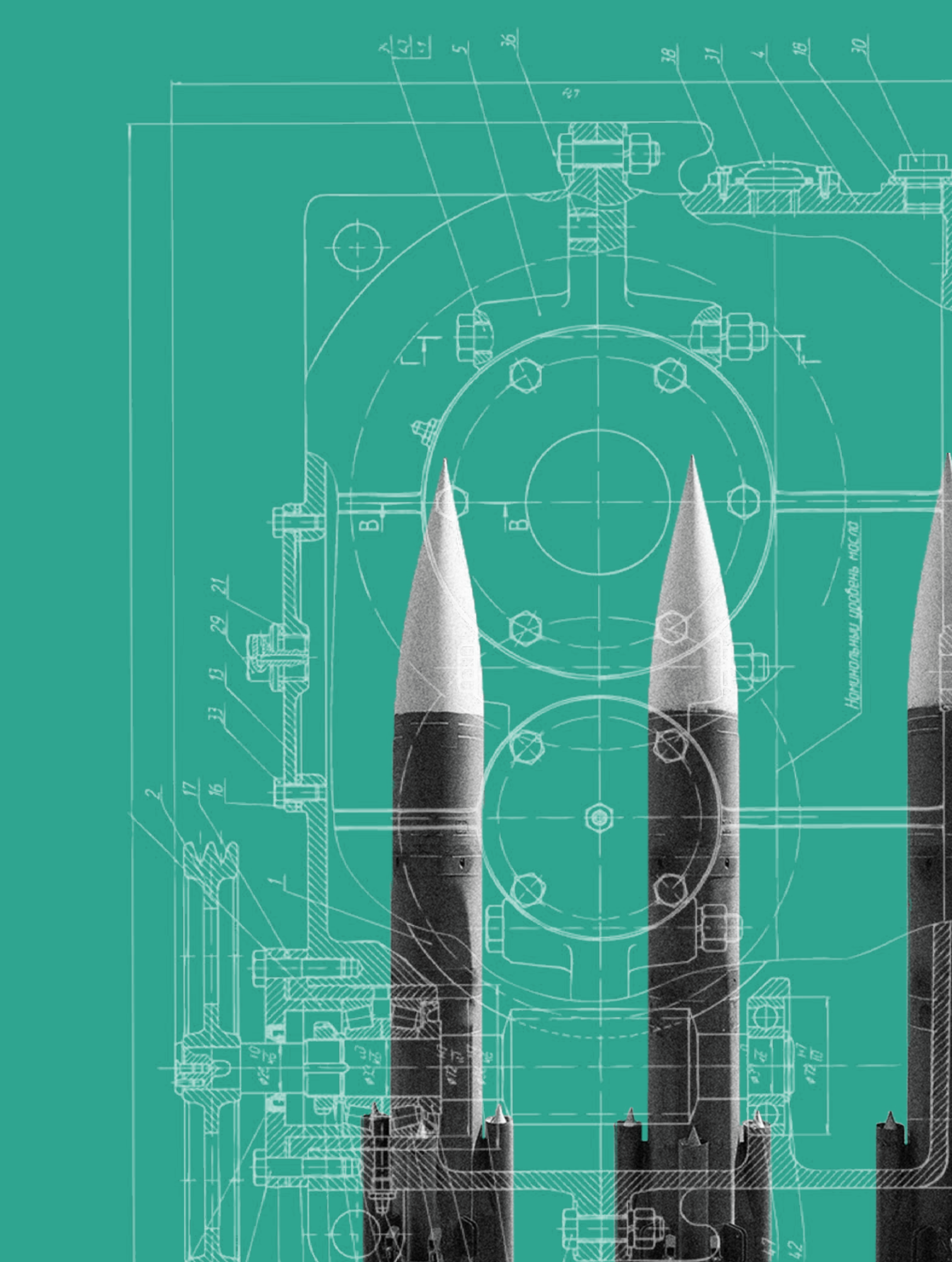
Taylor Owen

A podcast about technology's impact on our democracy, economy and society

www.bigtechpodcast.com

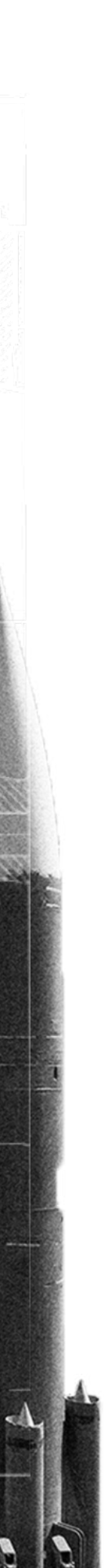
SUBSCRIBE:

Listen on Apple Podcasts / Spotify / Listen on Google Podcasts


Номинальный уровень масла

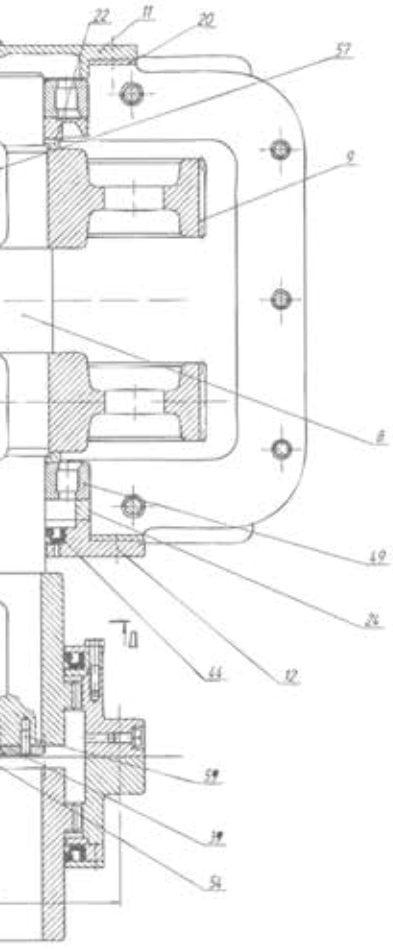
A New Arms Race and Global Stability

Amandeep Singh Gill



Two capabilities excite military planners more than anything else: first, the ability to get inside the decision-making loop of an adversary, and stay a step ahead of their responses in a conflict; second, the ability to sense the battle space fully and see what is going on at any place at any time (Hitchens 2019). And nothing makes military planners more jittery than the prospect of body bags, hence the attraction of finding ways to wage "bloodless" wars or do violence at a distance.

Artificial intelligence (AI) systems promise them a threefold bounty. First, data from different domains can be fused and weighted in real time by weapons platforms and related decision-support systems. Force can thus be applied at a faster tempo. Second, human limitations in digesting sensor inputs, say, live video feeds from a variety of locations, can be overcome for enhanced situational awareness, which could enable tailored and timely application of force for maximum effect. Finally, and perhaps not today but soon, machines can step into front-line combat roles, which mitigates the political implications of human casualties.



Of course, there are a number of important hurdles to be overcome first. The technology is still not mature. Very few AI models today can capture exceptions beyond the classifiers that are used during their training with data, nor can they learn in real time beyond such training on an incremental basis. There are safety issues; for example, image recognition models can be easily spoofed by so-called "adversarial" attacks (Vincent 2017). Even if an algorithm performs well in simulations, trust in the human-AI interface is still to be established, including in domains such as autonomous cars that have had years of testing and billions in investments. Commanders who have seen sailors or aviators struggle with simple digital dashboards would be loath to trust obscure and inscrutable algorithmic models. Not least, significant legal and ethical challenges remain to the ceding of human control and judgment to black-box algorithms.

At this stage of development, a useful analogy is the financial sector, where the stakes (and potential rewards) are high, as is the interconnectedness of risk, often in ways that are not so explicit or clear. A report by the Basel, Switzerland-based Financial Stability Board (FSB) highlights how "third-party dependencies" could grow as AI is introduced in financial services and how "new and unexpected forms of interconnectedness" could arise, for instance, as previously unrelated data sources come together (FSB 2017, 1).

Traditionally, arms control experts have looked at the introduction of new technologies of warfare from the perspectives of both stability and compliance with existing legal norms. In the latter context, there is no specific injunction against AI systems in existing arms control treaties on conventional weapons or weapons of mass destruction (nuclear, biological and chemical). How potential AI systems could impact existing international humanitarian law (IHL) principles such as distinction, proportionality and military necessity has been the subject of discussions in Geneva, Switzerland, since 2014 on "lethal autonomous weapons systems" under the 1980 Convention on Certain Conventional Weapons (CCW) (Gill, forthcoming 2020). A number of options to regulate such systems have been proposed to address concerns related to the undermining of IHL, in particular the fudging of human accountability for the laws of armed conflict. Moral and human rights concerns have also been cited to propose a complete ban on systems that can take life and death decisions without human intervention (Guterres 2018).

The stability arguments highlight several risks: a lowered threshold for use of lethal force; a new arms race (*arms race stability*); miscalculation and escalation due to the use of autonomous systems during tense faceoffs (*crisis stability*); and an undermining of the fragile balance in strategic weapons (*deterrence stability*) due to an AI-driven breakthrough in strategic offence or defence. During the Cold War, the two superpowers prohibited the deployment of nationwide missile defence systems through a bilateral treaty, since such technologies could have created an illusion of invulnerability and tempted one side to launch pre-emptive nuclear strikes on the other (Korda and Kristensen 2019). They invested in invulnerable systems, such as submarines carrying strategic missiles, to shore up deterrence. Today, AI could make it easier to follow submarines with nuclear weapons as they leave port for their deterrence patrolling stations, thereby allowing an adversary to neutralize what has been considered thus far the invulnerable leg of the nuclear triad. Further, an outlier could introduce immature AI technologies into highly destructive conventional or nuclear arms as it seeks to restore deterrence with a more powerful adversary or nudge it back to the negotiating table with outlandish systems. (See, for instance, the discussion on Russia's Poseidon underwater autonomous nuclear delivery system in Boulanin [2019].)

Despite the headlines and the catchy titles, the nature and the extent of the AI arms race are hard to discern at this stage. In many ways, the battlefield is still techno-commercial (Lee 2018). What is worrisome is that the AI technological rivalry among the major powers is coming at a time when mutual trust is low and when traditional structures for dialogue on arms control are withering away. Equally, because there are no dramatic markers of progress on the outside, unlike the Cold War experience of nuclear tests and missile launches, and because AI algorithms and their training data sets are inherently opaque, each side is guessing what the other is up to and probably attributing more AI military intent and capability than is necessitated. The upheaval and economic losses created by

the coronavirus disease 2019 pandemic have added to the uncertainty and mutual suspicion. The psychological backdrop for an arms race dynamic is very much in place.

Where are we likely to see this arms race play out first, and does it have the potential to become a global arms race, as has been the case with nuclear arms?

AI systems for perimeter defence of naval task forces, anti-submarine warfare, mine-detection and counter-mine measures, aerial defence against drones at sea, and seabed-deployed sensor networks, as well as submersibles for protecting communications cables, could see investments and eventual deployments by advanced navies. Investments in autonomous aerial combat vehicles, autonomous swarms, and target detection and acquisition systems, for force application from air and the navigation and control of supersonic and hypersonic combat systems, are likely to grow as well. On the ground, a range of logistics and support functions, as well as over-the-horizon reconnaissance and attack capabilities against high-value targets, are likely to see investments. AI use in some dirty and dangerous jobs, such as counterterrorism or IED (improvised explosive device) clearing operations, would also grow. In all these areas, since the relative quality of AI systems will be harder to assess than physically embodied weaponry, contestants will be left less certain as to each other's capabilities, increasing the risk of miscalculation and disproportionate responses in capability development and deployment.

A significant area of AI use is likely to be cyberwarfare capabilities. Today, cyberweapons do not mutate during use, but tomorrow they could do so autonomously in response to defensive measures. This could endow them with strategic effects.

Hopefully, strategic systems themselves will not see early and consequential integration of the AI technologies that are available today. This hope rests mainly on the culture of strategic communities, which prefer hard-wired systems with calculable certainties and failure rates. The risk of "entanglement" will remain, nonetheless, as new systems bring new types of data, actors and domains into the calculations of the strategic communities (for a pessimistic view, see Johnson [2020]). There will be pressure also on the offence-defence equation if there are AI breakthroughs in areas such as submarine detection and communications with, or control of, hypersonics. Another concern is the perceptions of parity among and between the players; some nuclear armed states may get an early-mover advantage by using AI to better manage the conventional and sub-conventional parts of the conflict escalation ladder, which might force others to threaten to increase their reliance on early nuclear weapons use or risky deployment postures.

In terms of geographical theatres of contestation, AI systems are likely to be deployed earlier in the maritime domain and in areas such as the North Atlantic/Arctic, the Gulf and the South China Sea in the Indo-Pacific. This is because of the sensitivity attached to shifts in balance of power in these areas and the operational focus of the major military powers.

A significant area of AI use is likely to be cyberwarfare capabilities.

Do AI weapons systems have the potential to impact the global balance of power? Possibly — but not so much as a stand-apart variable different from other trends driving shifts in power today. In that sense, relying on the historical experience with nuclear weapons can only take us so far with regard to AI systems. Proliferation could still turn the AI arms race among the major powers into a global phenomenon, and regional AI competition could throw up a few nasty deployment surprises. But the global balance of power will be shaped by many interdependent factors; digital technologies will be just one among many.

To conclude, what should be the key areas of immediate action for the international community to prevent the AI arms race from going global and to manage its international security consequences?

First, bring autonomous weapons systems into the agendas of current dialogues on disarmament, arms control and non-proliferation. Doing so would enhance transparency and encourage better understanding of intentions, capabilities and, eventually, deployment doctrines. It would also encourage sharing of best and "worst" practices, just as shared learning on safety and security of nuclear weapons was built up during the Cold War.

Second, discourage the commingling of strategic systems and AI-based decision-support systems. This work could take the form of political understandings among the nuclear-armed states. Additional understandings could be built around AI use that might impinge on the offence-defence equation.

Third, pursue discussions that have been taking place in Geneva among the United Nations' Group of Governmental Experts (2018) working in this area, to reach agreement on national mechanisms to review autonomous weapons systems with regard to obligations under IHL, and to exclude those systems that cannot comply with such obligations. Such an agreement could be accompanied by regular exchange of experience on the quality of the human-machine interface. Thus, use scenarios, where the pace of action on the battlefield exceeds the limits of human decision makers to exercise effective supervision or correctly interpret the information that AI systems are relaying to them, could be identified and avoided.

Decades ago, Jonathan Schell highlighted the danger that hair-trigger alert systems pose and argued powerfully for abolishing nuclear weapons (Schell 1998). Today, we need to advocate similarly for the "gift of time" in regard to autonomous weapons. After all, when we do something as quintessentially human as taking a deep breath, we allow wisdom to flow into the moment and enhance the quality of our actions.

WORKS CITED

Boulanin, Vincent, ed. 2019. *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk. Volume I: Euro-Atlantic Perspectives*. Stockholm, Sweden: Stockholm International Peace Research Institute.

FSB. 2017. *Artificial intelligence and machine learning in financial services: Market developments and financial stability implications*. Basel, Switzerland: FSB. www.fsb.org/wp-content/uploads/P011117.pdf.

Gill, Amandeep S. Forthcoming 2020. "The changing role of multilateral forums in regulating conflict in the digital age." *International Review of the Red Cross*.

Group of Governmental Experts. 2018. *Report of the 2018 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems*. CCW/GGE.1/2018/3. October 23.

Guterres, António. 2018. "United Nations Secretary-General — Remarks at 'Web Summit.'" Speech, Lisbon, Portugal, November 5. www.un.org/sg/en/content/sg/speeches/2018-11-05/remarks-web-summit.

Hitchens, Theresa. 2019. "Navy, Air Force Chiefs Agree To Work On All Domain C2." *Breaking Defense*, November 12. https://breakingdefense.com/2019/11/exclusive-navy-air-force-chiefs-agree-to-work-on-all-domain-c2/.

Johnson, James S. 2020. "Artificial Intelligence: A Threat to Strategic Stability." *Strategic Studies* Quarterly 14 (1): 16–39. www.airuniversity.af.edu/Portals/10/SSQ/documents/Volume-14_Issue-1/Johnson.pdf.

Korda, Matt and Hans M. Kristensen. 2019. "US ballistic missile defenses, 2019." *Bulletin of the Atomic Scientists* 75 (6): 295–306.

Lee, Kai-Fu. 2018. *AI Superpowers: China, Silicon Valley, and the New World Order*. Boston, MA: Houghton Mifflin Harcourt.

Schell, Jonathan. 1998. *The Gift of Time: The Case for Abolishing Nuclear Weapons Now*. New York, NY: Metropolitan Books.

Vincent, James. 2017. "Google's AI thinks this turtle looks like a gun, which is a problem." The Verge, November 2. www.theverge.com/2017/11/2/16597276/google-ai-image-attacks-adversarial-turtle-rifle-3d-printed.

ABOUT THE AUTHOR

Amandeep Singh Gill is the director of the International Digital Health & AI Research Collaborative project at the Graduate Institute of International and Development Studies in Geneva, Switzerland. As an Indian foreign service officer, Amandeep served abroad in Tehran, Colombo and Geneva, and at Indian Foreign Service headquarters covering bilateral and multilateral issues related to political affairs, security, non-proliferation, technology, development and human rights. He was head of the Disarmament and International Security Affairs Division, Ministry of External Affairs, from 2013 to 2016, and India's ambassador and permanent representative to the Conference on Disarmament in Geneva, from 2017 to 2018. In 2017, Amandeep helped set up the Task Force on Artificial Intelligence for India's Economic Transformation and chaired the Group of Governmental Experts of the Convention on Certain Conventional Weapons on emerging technologies in the area of lethal autonomous weapon systems, which agreed on a set of guiding principles in 2018. As executive director, Secretariat of the UN Secretary-General's High-level Panel on Digital Cooperation, Amandeep helped the chairs, Melinda Gates and Jack Ma, deliver a path-breaking report in June 2019. Amandeep has a B.Tech. in electronics and electrical communications, Panjab University, Chandigarh; an advanced diploma in French history and language, Geneva University; and a Ph.D. in nuclear learning in multilateral forums, King's College, London.

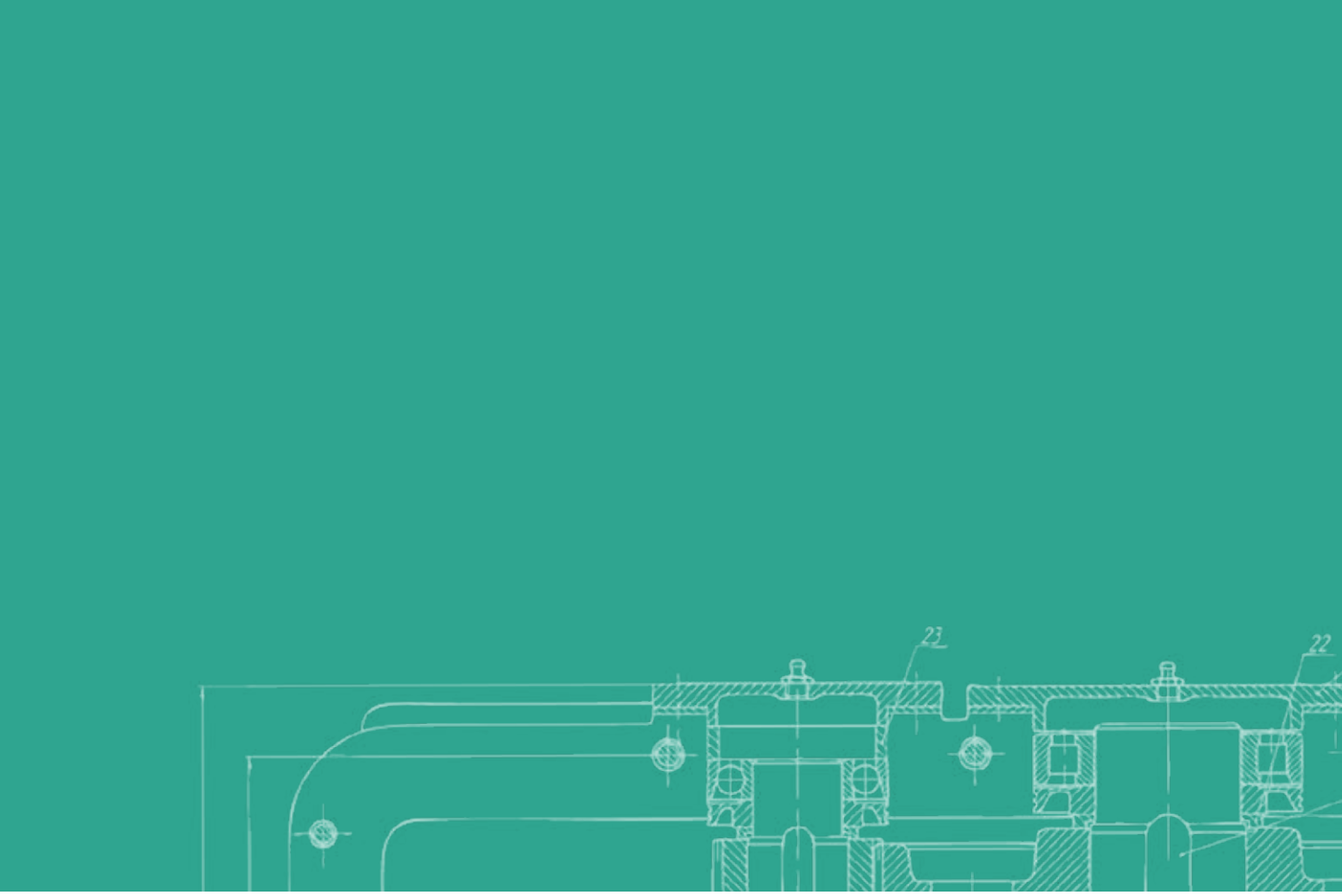
Security, Intelligence and the Global Health Crisis

A CIGI Essay Series

The impact of COVID-19 both globally and in Canada has raised important questions about best practices with regard to global and domestic health surveillance, early warning and preparedness. What role do security and intelligence institutions play in protecting societies against pandemic outbreaks?

• • •

cigionline.org/security-and-health



Public and Private Dimensions of AI Technology and Security

Maya Medeiros

The rapid emergence of disruptive technologies such as artificial intelligence (AI) requires that new governance frameworks be created so that these technologies' development can occur within a secure and ethical setting, to both mitigate their risks and maximize their benefits for humanity. There are public and private dimensions to AI governance. Various private companies have called for increased public regulation to ensure the ethical use of new technology, and some have even suspended high-risk applications, such as facial recognition for law enforcement, until a proper regulatory framework is in place. Public-private collaboration is essential to creating innovative governance solutions that can be adapted as the technology develops, not only to support innovation and commercial application but also to provide sturdy guardrails that protect human rights and social values.

Private Initiatives in AI Governance

Private companies' governance initiatives generally involve best practices and voluntary guidelines to govern the development and use of responsible AI.

An initial report on AI governance was launched in 2016 by the Institute of Electrical and Electronics Engineers (IEEE). The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems aims to "ensure every stakeholder involved in the design and development of autonomous and intelligent systems is educated, trained, and empowered to prioritize ethical considerations so that these technologies are advanced for the benefit of humanity" (IEEE 2017, 3). The initiative also involves a series of voluntary IEEE standards that address governance and ethical aspects of AI.

Another private initiative, the Partnership on AI, was established by several large technology companies — Apple, Amazon, DeepMind and Google, Facebook, IBM and Microsoft — and has since expanded to include a wide variety of companies, think tanks, academic AI organizations, professional societies, and charitable groups such as the American Civil Liberties Union, Amnesty International, the United Nations' Children's Fund and Human Rights Watch. The partnership's work involves study, discussion, identification, sharing and recommendation of best practices in the research, development, testing and fielding of AI technologies. The partnership addresses such areas as fairness and inclusivity, explanation and transparency, security and privacy, values and ethics, collaboration between people and AI systems, interoperability of systems, and the trustworthiness, reliability, containment, safety and robustness of the technology.

Many large technology companies have value-based principles for internal AI activities to guide their conduct.

The Information Technology Industry Council (ITI), a trade association, has developed its own set of AI principles for designing AI technologies beyond compliance with existing laws (ITI 2017). The ITI recognizes the potential uses and misuses of technology, the implications of its use or misuse, and the industry's responsibility and opportunity to take steps to avoid the reasonably predictable misuse of AI by committing to ethics by design.

Many large technology companies have value-based principles for internal AI activities to guide their conduct.[1] However, these private initiatives are not binding and require voluntary compliance by companies using the technology.

Public Initiatives in AI Governance

Public governance initiatives include international value-based policies for responsible AI and guidance for national legislation.

The Organisation for Economic Co-operation and Development (OECD) developed principles on AI to promote trustworthy AI that respects human rights and democratic values. The "OECD AI Principles," formally known as the *Recommendation of the Council on Artificial Intelligence*, were adopted in May 2019 by OECD member countries and are the first such principles signed on to by governments (OECD 2019a). Beyond OECD members, other countries, including Argentina, Brazil, Costa Rica, Malta, Peru, Romania and Ukraine, have already adhered to the OECD AI Principles, with further adherents anticipated. The OECD AI Principles set standards for AI that complement existing OECD standards in areas such as privacy, digital security risk management and responsible business conduct. The principles identify five complementary value-based principles for the responsible stewardship of trustworthy AI and also provide five recommendations to governments. While not legally binding, they aim to set the international standard for responsible AI and to help governments design national legislation. In June 2019, the Group of Twenty (G20) adopted human-centred AI principles that draw on the OECD AI Principles, which affirmed at the G20 level "that the AI we want is centered on people, respects ethical and

democratic values, [and] is transparent, safe and accountable" (OECD 2019b).

In addition to international principles, multiple foreign governments have presented national AI policies or policies that purport to regulate some aspect of the adjacent technology stack. In Canada, the National Cyber Security Strategy presents a vision for protecting Canadians' digital privacy, security and economy and a commitment to collaborate with France on ethical AI (Public Safety Canada 2018).

China has a national recommended standard for personal data collection, issued as GB/T 35273-2020 or "Information Security Technology — Personal Information Security Specification" (People's Republic of China 2020), which addresses data considerations similar to those in the European Union's General Data Protection Regulation. China's "Next Generation Artificial Intelligence Development Plan" highlights the need to strengthen research and establish laws, regulations and ethical frameworks on legal, ethical and social issues related to AI and protection of privacy and property (People's Republic of China 2017). In India, there is discussion on the importance of AI ethics, privacy, security and transparency, as well as on the current lack of regulations around privacy and security (National Institute for Transforming India 2018).

The European Union's Committee on Legal Affairs recommends that "the existing Union legal framework should be updated and complemented, where appropriate, by guiding ethical principles in line with the complexity of robotics and its many social, medical and bioethical implications" (European Parliament 2017, 9). The European Commission (2018a) published its strategy paper on AI but did not propose any new regulatory measures for AI. As a follow-up, it published a "Coordinated Action Plan on AI" that set forth its objectives and plans for an EU-wide strategy on AI (European Commission 2018b). A UK strategy considers the economic, ethical and social implications of advances in AI and recommends preparing for disruptions to the labour market, open data and data protection legislation, data portability, and data trusts. The UK perspective is centred on the fact that large companies that have control over vast quantities of data must be prevented from becoming overly powerful. France aims to implement inclusive and diverse AI and avoid the "opaque privatization of AI or its potentially despotic usage" (Macron, quoted in Rabesandratana 2018).

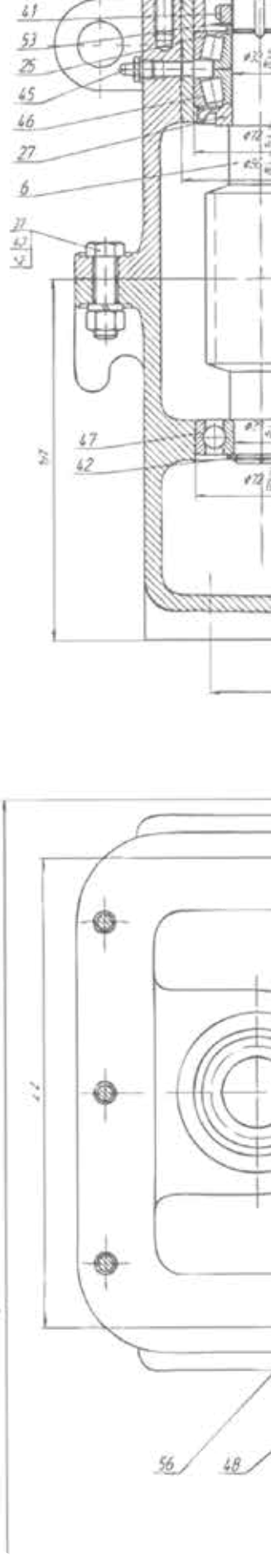
The United States appears to be focused on the military aspects of AI policy, with the House Committee on Armed Services legislating a National Security Commission on Artificial Intelligence mandated "to consider the methods and means necessary to advance the development of artificial intelligence, machine learning, and associated technologies by the United States to comprehensively address the national security and defense needs of the United States."[2] The commission's latest report recommends a White House-led technology council and aims to convey one big idea: "The countries, companies, and researchers that win the AI competition — in computing, data, and talent — will be positioned to win a much larger game" (National Security Commission on Artificial Intelligence 2020, 2).

However, these policies are not binding on private players. Governments are slowly starting to introduce binding national laws directed to certain technologies, such as automated decision making, face recognition and conversational agents. While technology development and deployment accelerates, private actors continue to request increased regulation to ensure the ethical use of new technology and mitigate risk.

The Need for a Joint Effort

A joint effort by private companies and public governments is needed to create a more agile regulatory framework responsive to the accelerating pace of disruptive technologies. Many private entities better comprehend the AI tools and unintended impacts of regulation, making their perspectives essential to public regulators.

Public action is required to mandate compliance with AI policy and enforce ethical requirements. There should be coordinated effort between different public actors. Different regulators with similar policy objectives should adopt universal language for legislation to encourage regulatory convergence. International standards with universal language can help streamline adoption by private players operating in multiple countries,



Innovative regulatory models for disruptive technologies are also emerging.

for example. Private actor cooperation is required for widespread compliance with any new regulation.

Current AI policies are often value-based and might not provide enough detail on how private actors can achieve the target objectives within different use cases in order to comply with the policy. There should be sufficient guidance to understand how a specific AI tool should be designed to meet an objective and whether the specific tool is compliant with the objective. Any regulation should include operational guidance and workable directives developed in cooperation with private actors. Detailed examples for different applications can help to provide guidance and create more certainty on a specific policy's impact on those applications.

Private players can agree to voluntarily adopt public initiatives until such time as new governance solutions are available. If a company voluntarily adopts an AI policy, it must then comply with the policy. Evaluating specific AI tools for compliance with principles requires the dedication of significant efforts on the part of the private player. A comprehensive evaluation often requires a technical understanding of a specific AI tool to see how it maps to different principles. However, there can be a lack of knowledge by regulators and a need for input by private industry to improve understanding of these complex technologies. Enforcement of a policy might require examination of the AI tool, which is undesirable if aspects of the tool are protected as trade secrets. Protective measures for the code will be required for examinations of code mandated by policy.

Private contracts can be used to increase adoption of governance terms for new disruptive technologies. Ethics and governance requirements that might otherwise be voluntary can be incorporated in contracts to create binding obligations between private players. However, incorporating governance terms into contracts requires agreement by the contracting parties. If parties were to commonly include governance terms in contracts relating to AI technology, they would help to encourage the adoption and standardization of these terms and to establish at least minimum standards for ethics and security.

Open-source software licences can also be used to encourage adoption of governance terms. Disruptive technologies are commonly being offered as "open-source" tools licensed by standard terms. The software licences could also be updated to include minimum governance requirements, such as through the listing of both permitted ethical uses of the open source tools *and* prohibited uses. Widespread use of the tools could in turn trigger widespread adoption of these minimum governance terms. For example, contact-tracing tools to track outbreaks for public health purposes could involve collecting data with varying levels of sensitivities. Contact-tracing tools and associated data can be released under terms of use that mandate basic ethical practices.

Innovative regulatory models for disruptive technologies are also emerging. New hybrid "regulatory markets" pair strong government oversight with private sector regulators. In these regulatory markets, private sector regulators compete for the right to regulate specific AI fields. Instead of enacting traditional regulation, government can set the goals, and independent companies can determine how they should meet such goals, thereby incentivized to invent streamlined ways to achieve these government-set goals. There are risks that private regulators will be influenced by the entities they regulate rather than by public interest, and it will be important to ensure that private regulators act independently.

Private actors will continue to request public regulatory guidance for high-risk applications of disruptive technology, such as the use of AI in self-driving cars and law enforcement. Until such time as proper security and ethical regulatory measures are in place, the use of these technologies will be stifled. However, uncertainties around regulatory compliance and enforcement can also stifle innovation. We need new solutions for regulating

disruptive technology that are responsive to high-risk applications while also supporting technology development and deployment.

NOTES

1 See www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1:primaryr6; Pinchai (2018); IBM (2018).

2 *John S. McCain National Defense Authorization Act for Fiscal Year 2019*, Pub L No 115-232, §1051(b)(1), 132 Stat 1636 at 1964.

WORKS CITED

European Commission. 2018a. "Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe." COM(2018) 237 final, April 25. Brussels, Belgium: European Commission. https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe.

———. 2018b. "Coordinated Plan on Artificial Intelligence." COM(2018) 795 final, December 7. Brussels, Belgium: European Commission. https://ec.europa.eu/knowledge4policy/ai-watch/coordinated-action-plan-ai_en.

European Parliament. 2017. *Report with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))*. A8-0005/2017. January 27. www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.pdf.

IBM. 2018. "IBM's Principles for Trust and Transparency." *THINKPolicy Blog*, May 30. www.ibm.com/blogs/policy/trust-principles/.

IEEE. 2017. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. Version 2 — For Public Discussion*. December. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf.

ITI. 2017. "AI Policy Principles." October 24. Washington, DC: ITI. www.itic.org/resources/AI-Policy-Principles-FullReport2.pdf.

National Institute for Transforming India. 2018. "National Strategy for Artificial Intelligence." Discussion Paper, June. https://niti.gov.in/writereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf.

National Security Commission on Artificial Intelligence. 2020. *National Security Commission on Artificial Intelligence 2020 Interim Report and Third Quarter Recommendations*. October. https://drive.google.com/file/d/1R5XqQ-8Xg-b6CGWcaOPPUKoJm4GzjpMs/view.

OECD. 2019a. *Recommendation of the Council on Artificial Intelligence*. OECD/LEGAL?0449. Adopted on May 21. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

———. 2019b. "Remarks by Angel Gurría, OECD Secretary-General, Osaka, Japan." 2019 G20 Leaders' Summit — Digital (AI, data governance, digital trade, taxation), June 28. www.oecd.org/about/secretary-general/2019-g20-leaders-summit-digital-osaka-june-2019.htm.

People's Republic of China. 2017. "Next Generation Artificial Intelligence Development Plan issued by State Council." China Science and Technology Newsletter No. 17, September 15. Beijing, China: Department of International Cooperation Ministry of Science and Technology. http://fi.china-embassy.org/eng/kxjs/P020171025789108009001.pdf.

———. 2020. [Information Security Technology — Personal Information Security Specification.] GB/T 35273-2020. Issued by General Administration of Quality Supervision, Inspection and Quarantine and State Administration for Market Regulation on March 6; implemented on October 1. www.secrss.com/articles/17713.

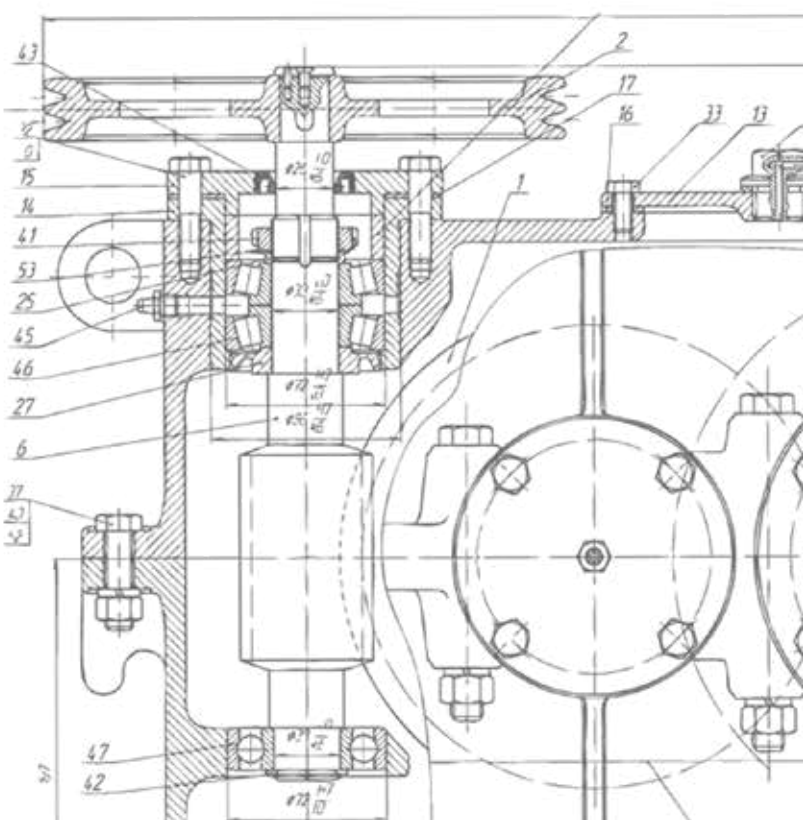
Pinchai, Sundar. 2018. "AI at Google: our principles." *The Keyword* (blog), June 7. www.blog.google/technology/ai/ai-principles/.

Public Safety Canada. 2018. *Canada's Vision for Security and Prosperity in the Digital Age*. Cat. No. PS4-239/2018E. Ottawa, ON: Public Safety Canada. www.publicsafety.gc.ca/cnt/rsrcs/pblctns/ntnl-cbr-scrt-strtg/ntnl-cbr-scrt-strtg-en.pdf.

Rabesandratana, Tania. 2018. "Emmanuel Macron wants France to become a leader in AI and avoid 'dystopia.'" March 30. *Science*, March 30. www.sciencemag.org/news/2018/03/emmanuel-macron-wants-france-become-leader-ai-and-avoid-dystopia.

ABOUT THE AUTHOR

Maya Medeiros is an intellectual property (IP) lawyer, patent agent (Canada, United States) and trademark agent (Canada, United States), and has a degree in mathematics and computer science. She has extensive technical experience in AI, blockchain, cybersecurity, cryptography, payments, virtual and mixed reality, wearables, and other computer-related technologies. She is a key contributor to www.aitech.law on the ethical and legal implications of AI. She advises on IP strategy and develops programs to build global IP portfolios. She prepares benchmark IP analysis, involving the technical review of patent portfolios, competitor research and patent litigation. Maya is a partner at Norton Rose Fulbright Canada LLP.

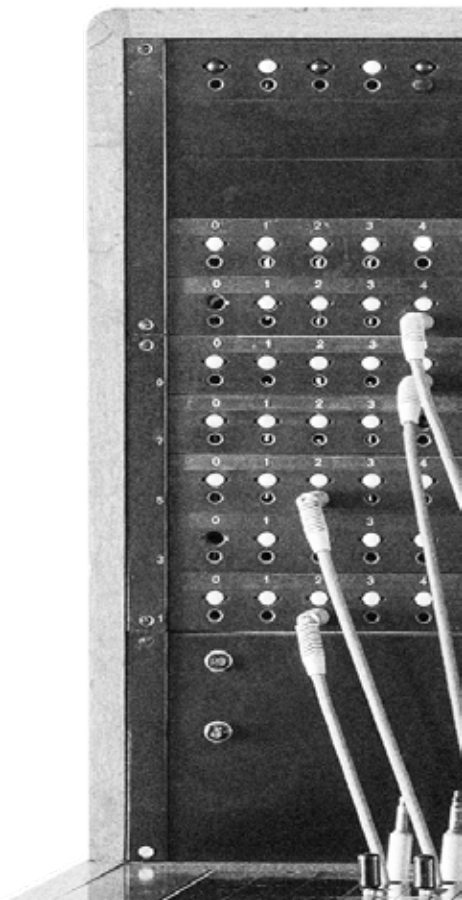


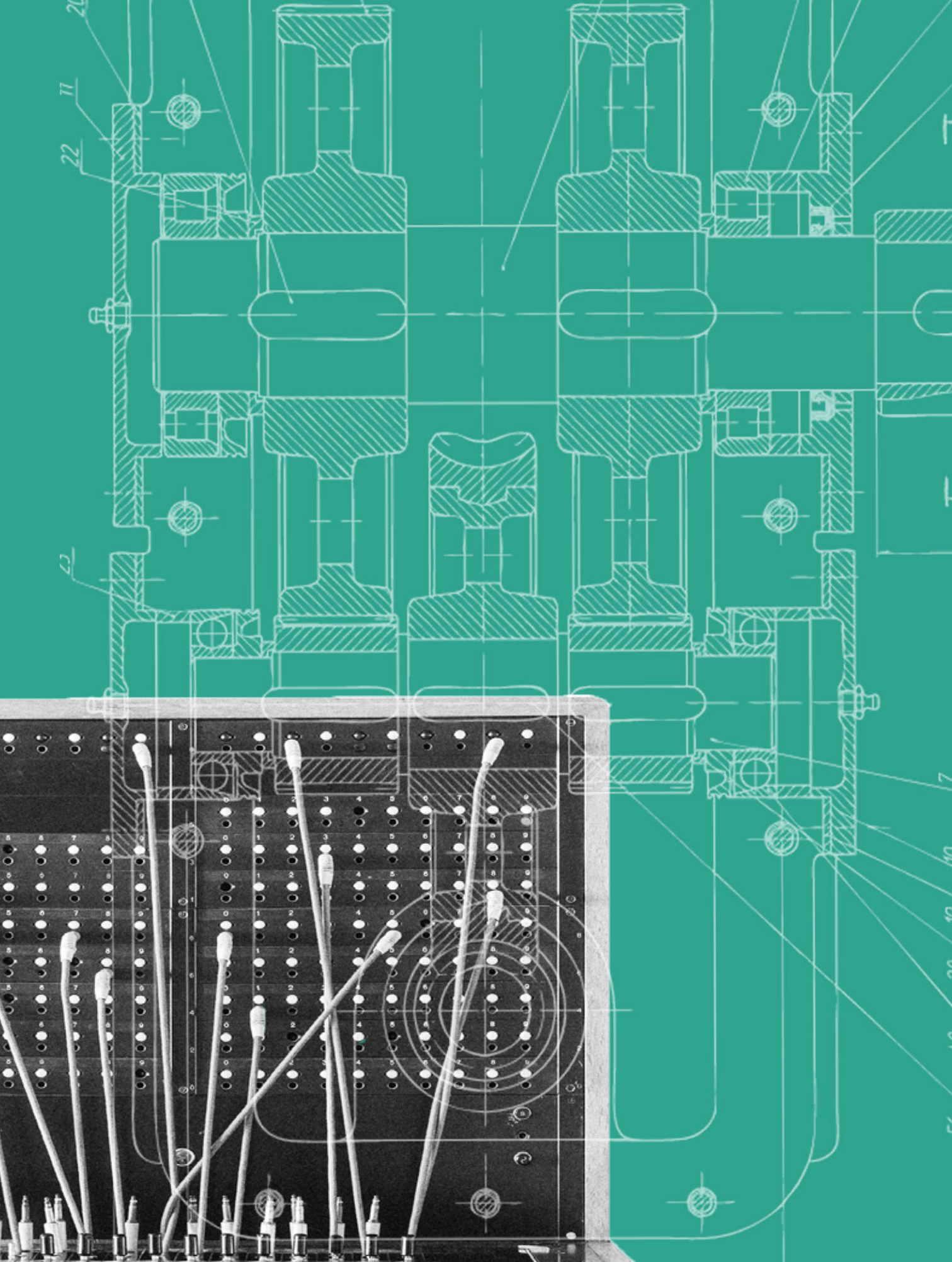
International Legal Regulation of Autonomous Technologies

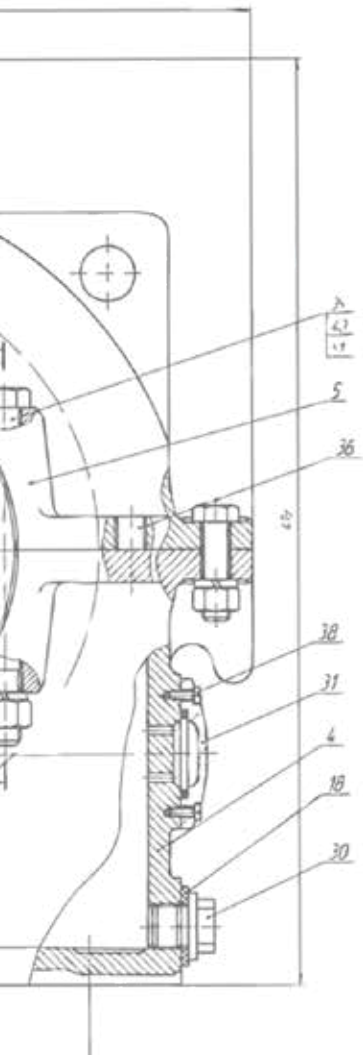
Liis Vihul

The advent of new technologies always prompts questions concerning their legality, and this is certainly true with respect to autonomous technologies, including those using varying degrees of artificial intelligence (AI). As autonomous solutions are developed and employed, countries need to ensure that their use aligns with established moral and ethical principles, which are often enshrined in both domestic and international legislation. The basic legal dilemma concerning any new technology is ascertaining whether existing law is capable of regulating it in conformity with those principles and, if not, what new legal instruments are necessary to meet that objective. This essay explores that question in the context of autonomous technologies.

Innovation in autonomy is being driven simultaneously by civilian and national security (including military) demands. Commercial autonomous technology for civilian application is primarily subject to domestic legal regulation, although international law can, and is likely to, play some role in its governance. Autonomous military technologies are predominantly developed for employment in an international environment during armed conflict; in that setting, international law is prominent. The essay begins with a discussion of the prospect for international legal regulation of autonomous civilian technologies. It then turns to certain challenges relating to international law that employing autonomy in national security and defence contexts, including on the battlefield, presents.







Regulation for Civilian Purposes

Legislatures across the globe should be preparing to amend their laws, and possibly adopt new ones, governing autonomous technologies. Some applications, such as aircraft autopilot systems and industrial robots, have been employed for decades, albeit in strictly controlled environments where robust security controls are in place. In the future, technologies with varying degrees of autonomy will become pervasive in many societies. Driverless public transit systems, self-driving cars and AI algorithms in medical diagnosis are leading this innovation, with countless other use cases bound to follow. Inevitably, domestic laws will require some degree of revision to ensure adequate regulation of such systems.

For the present, these new technologies are primarily subject to industry self-regulation, with several large companies having adopted internal policies relating to the use of automation in their products and services (for examples, see International Committee of the Red Cross 2019, 25–26). The experience states have had with current digital technologies offers valuable lessons in this regard; when the private sector is left to self-regulate, friction between companies and governments is likely to arise. Criticism by states directed at Twitter and Facebook about their handling of online content, such as fake news and live streaming of violent incidents, or the susceptibility of their algorithms to manipulation and biases, is illustrative. That these companies have called on governments to specify through regulation the kinds of action expected of them is therefore unsurprising (Press Association 2019; Rudgard and Cook 2019).

The extent to which industry self-regulation can govern more advanced autonomous technologies to the satisfaction of governments, civil society and the public generally is limited. Google, itself, has acknowledged that "self- and co-regulatory approaches will remain the most effective practical way to address and prevent AI related problems in the vast majority of instances, within the boundaries already set by sector-specific regulation," but that "there are some instances where additional rules would be of benefit" and "relying on companies alone to set standards is inappropriate" (Google, n.d., 29; Evans 2020). Accordingly, it is sensible for governments to engage with the private sector and collaboratively work toward optimal governance regimes, as opposed to intervening only when unwanted consequences of this new technology have begun to manifest.

Regulatory rules, rather than legislative solutions, are likely to emerge first, as has been the case with other novel technologies. In the field of nanotechnology, for example, several European countries have adopted regulations that impose reporting requirements on companies that manufacture, import or distribute nanomaterials.[1] In the field of autonomy, we can likewise expect regulations tackling discrete issues, which at some point will be followed by legislative action, whether through amendments to existing laws or the adoption of new ones (this is without prejudice to the adoption of so-called enabling legislation, that is, legislation that grants the power to adopt regulations to a certain person or entity, such as a government minister).

Public international law, by contrast, will largely play a bystander's role insofar as commercial autonomous solutions meant for civilian use are concerned. However, the international community may at some point feel the need to harmonize countries' domestic laws to ensure that the internal legal regulation of these commercial technologies is consistent across borders. The legal mechanism for harmonization would be the adoption of a so-called uniform law treaty that obligates states that are parties to the instrument to legislate domestically with respect to their criminal, civil or administrative laws. For example, such a treaty could prescribe uniform safety standards, liability rules, certification schemes, data management processes, human supervision requirements over the use of the technology, fail-safe mechanisms to be put in place, operational constraints, rules regarding bias, and criminal offences involving autonomous technologies.

The most likely starting point for international legal regulation along these lines would be the European Union, for it is the only international organization with an institutional capacity, a pre-existing mandate and the political appetite to adopt such far-reaching binding rules (in fact, it has already taken preliminary steps toward intra-community regulation of AI; see European Commission 2020). Although formally the European Union only has the authority to legislate

vis-à-vis its member states, the effect of any resulting regulation would extend beyond the organization's borders. The situation might be analogous to the European Union's General Data Protection Regulation — to the extent foreign companies offering products and services in the field of automation want to operate in the EU market, they would be obliged to follow applicable EU rules. This presents a strategic opportunity for the European Union, for it is uniquely well-positioned to serve as a pioneer in this area, thereby shaping the conversation as to the appropriate legal and regulatory regime for autonomous technologies.

Regulation of National Security and Defence-related Autonomous Technologies

It is widely accepted that autonomy, in particular AI, will revolutionize warfare. Examples of contexts in which autonomy is and will be employed include information processing, notably intelligence analysis; unmanned weapon systems; realistic military training; psychological warfare; and military command and control. It is therefore unsurprising that great-power competition for supremacy in military autonomous technologies and AI is under way.

In that warfare is governed by a dense international legal framework, many rules already exist that regulate the use of autonomous technologies in war. These rules form a regime of international law known as international humanitarian law (IHL), also labelled the law of armed conflict.

Scholarship on the interplay between these new technologies and IHL has primarily focused on the use of lethal autonomous weapons (Schmitt and Thurnher 2013; O'Connell 2014; Sassòli 2014; Geiss 2015). At the state level, a group of governmental experts convened under a UN umbrella has confirmed that "international humanitarian law continues to apply fully to all weapons systems, including the potential development and use of lethal autonomous weapons systems" (Group of Governmental Experts 2017, para. 16(b)), which logically leads to the conclusion that other military uses of autonomous technologies are likewise governed by this subfield of international law.

IHL, in particular its rules governing the conduct of hostilities (that is, the way in which a war is waged), are relevant insofar as the international community has not prohibited particular means or methods of warfare. Presently, no automated or autonomous technologies have been banned, although states have been under political, scholarly and civil society pressure to prohibit fully autonomous lethal weapons since the launch of the "Ban Killer Robots"[2] movement. For instance, the European Parliament in 2018 adopted a resolution in which it urged the European Commission, individual member states and the European Council to "work towards the start of international negotiations on a legally binding instrument prohibiting lethal autonomous weapon systems" (European Parliament 2018, para. 3). In the absence of such a treaty, existing IHL rules govern their use.

The issue of lethal autonomous weapon systems aside, it is clear that autonomous technologies will increasingly find military usage. It is equally clear that the application of the pre-automation, pre-autonomy rules of IHL to those technologies is not without challenges. Many existing debates over how to apply IHL rules would apply equally to autonomous systems, as in the case of questions concerning the permissibility of directing non-destructive military operations against civilian objects[3] or the geographical boundaries of the applicability of humanitarian law.[4]

It is widely accepted that autonomy, in particular AI, will revolutionize warfare.

Yet, issues unique to autonomy are bound to arise as well. For example, a cross-cutting issue in IHL, as well as in related fields of international law such as international criminal law, concerns accountability. If, for instance, autonomous cyber capabilities unexpectedly cause harm to civilians or damage to civilian objects, questions of responsibility attach. Under IHL, states are responsible for ensuring their weapon systems are used in a manner consistent with the conduct of hostilities rules. This obligation begs the question of weapon

systems that operate autonomously, perhaps even using AI to select targets. If the armed forces using a system cannot assess the harm likely to be caused to the civilian population or civilian objects by an autonomous system with the requisite degree of reliability, whatever the correct standard of likelihood is, those armed forces are using the weapon indiscriminately in the battlespace. This would constitute a breach of IHL by the state employing the autonomous weapon system.[5]

Furthermore, international criminal law imposes individual criminal responsibility for war crimes, which include directing attacks against civilian objects with "intent" and "knowledge."[6] Questions about how criminal tribunals would apply these notions to encompass civilian damage caused by autonomous systems in circumstances such as those mentioned above would loom large in any criminal prosecution.

A more practical challenge is that it is difficult to regulate something that one does not fully understand.

Autonomy is also being used for national security purposes, both benign and malicious, beyond the battlefield. As malicious uses are exposed, they often raise legal and ethical alarm bells. The highest-profile case of a government resorting to these technologies to surveil and identify individuals is the Chinese government's continuous monitoring of the Uighur Muslim minority (Taddonio 2019), a case that set a precedent for other authoritarian governments to employ advanced technologies for illicit purposes. Adding to the complexity of the situation is commercial opportunism. The case of Clearview AI — a facial recognition software company that automatically scrapes images from the internet to form a database of several billion files, thereby enabling facial recognition (Hill 2020) — is a telling example of how the private sector, if left to self-regulate, risks societal harm that is not necessarily outweighed by the legitimate use of their services for national security and public order purposes. These and other cases demonstrate the potential negative effects of autonomy and automation, including the erosion of human rights (such as the right to privacy, freedom of the press and freedom of assembly). They also highlight the need to pay even greater attention to preserving and safeguarding the rule of law and basic moral and ethical values in the face of technological developments.

Conclusion

New technologies present normative challenges to both domestic and international law, in particular with regard to the suitability of pre-existing rules. Certain technology-specific issues are inevitably bound to arise that will require regulatory and legislative action. The resulting normative evolution will first occur in the domestic setting, for international law making is a relatively slow process, especially in fields with a national security nexus.

In this process, states will face many challenges. A fundamental difficulty stems from the dual-use nature of autonomous solutions. Accordingly, both domestic regulators and legislatures, as well as states as they engage in the interpretation and adoption of international law, will need to tread carefully, ensuring, on the one hand, that the rules and interpretive positions they adopt do not stifle innovation while guaranteeing, on the other, that they effectively prevent malicious uses of the technology. Sensible normative frameworks must be collaborative; governments should therefore work with industry and civil society in adopting fit-for-purpose governance regimes, while states should work together to fashion rules that advance shared values.

A more practical challenge is that it is difficult to regulate something that one does not fully understand. Autonomous technologies are in their infancy, and predicting scientific developments in this field — even in the near term — is difficult, if not impossible. Any new laws and regulations will need to be sufficiently general so as to not become outdated quickly, but also not so vague that they provide no meaningful guidance. The difficulty of this undertaking inevitably means that no overarching area-specific rule set will be adopted in the near future — neither domestic

legal acts governing autonomous technologies writ large, nor an international treaty on autonomy as such. Instead, we may expect discrete rules governing relatively specific aspects of autonomous technologies.

NOTES

1 See, for example, European Union Observatory for Nanomaterials, National Reporting Schemes, https://euon.echa.europa.eu/national-reporting-schemes.

2 For more information, see www.stopkillerrobots.org/.

3 The principle of distinction is set forth in, *inter alia*, Article 48 of Additional Protocol I to the Geneva Conventions (Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts, 8 June 1977, 1125 UNTS 3): "In order to ensure respect for and protection of the civilian population and civilian objects, the Parties to the conflict shall at all times distinguish between the civilian population and combatants and between civilian objects and military objectives and accordingly shall direct their operations only against military objectives." The term "military operations" is generally understood as prohibiting the parties' "attacks" in the sense of Article 49(1) of Additional Protocol I against the civilian population, individual civilians and civilian objects. A degree of uncertainty exists as to whether military operations that do not amount to "attacks" can also be considered in violation of the principle of distinction.

4 For instance, whether IHL's applicability would extend to the territories of other (non-adjacent) states in an armed conflict between government armed forces and an organized armed group is a matter of controversy. If, for example, a member of the organized armed group travelled to an overseas country and launched a destructive autonomous cyberspace operation against the state that they are fighting, the issue arises as to whether that person and the information technology (IT) equipment that the person is using are subject to IHL. Some experts are of the view that in such a circumstance IHL continues to apply vis-à-vis that person and the equipment, in which case killing that person, and damaging or destroying the equipment (for example, by way of remote cyber operations), would not constitute a breach of IHL. This is because members of organized armed groups, as well as any objects qualifying as military objectives (in this case, the IT equipment), are targetable during an armed conflict. Others posit that IHL does not follow a person and objects in said manner and that the situation would instead be governed by international human rights law. It should also be noted that a scenario of this type involves other complex legal issues, for instance, the legal basis for the state that is engaging lethal or destructive operation in another state's territory.

5 Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts, 8 June 1977, 1125 UNTS 3, Art. 51(4)(a).

6 Rome Statute of the International Criminal Court, 17 July 1998, 2187 UNTS 90, Arts. 8(b)(2) and 30.

WORKS CITED

European Commission. 2020. "On Artificial Intelligence — A European approach to excellence and trust." COM(2020) 65 final, February 19. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

European Parliament. 2018. *Resolution on autonomous weapon systems*. 2018/2752(RSP), September 12.

Evans, Zachary. 2020. "Google CEO Calls for Government Regulation of Artificial Intelligence." *National Review*, January 20. www.nationalreview.com/news/google-ceo-calls-for-government-regulation-of-artificial-intelligence/.

Geiss, Robin. 2015. "The International-Law Dimension of Autonomous Weapons Systems." International Policy Analysis. Berlin, Germany: Friedrich-Ebert-Stiftung.

Google. n.d. "Perspectives on Issues in AI Governance." White paper. https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf.

Group of Governmental Experts. 2017. *Report of the 2017 Group of Governmental Experts on Lethal Autonomous Weapons Systems*. UN Doc CCW/GGE.1/2017/3. December 22. https://undocs.org/CCW/GGE.1/2017/3.

Hill, Kashmir. 2020. "The Secretive Company That Might End Privacy as We Know It." *The New York Times*, January 18. www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html.

International Committee of the Red Cross. 2019. "Autonomy, artificial intelligence and robotics: Technical aspects of human control." August. Geneva, Switzerland: International Committee of the Red Cross. www.icrc.org/en/document/autonomy-artificial-intelligence-and-robotics-technical-aspects-human-control.

O'Connell, Mary Ellen. 2014. "Banning Autonomous Killing: The Legal and Ethical Requirement That Humans Make Near-Time Lethal Decisions." In *The American Way of Bombing: Changing Ethical and Legal Norms, from Flying Fortresses to Drones*, edited by Matthew Evangelista and Henry Shue, 224–36. Ithaca, NY: Cornell University Press.

Press Association. 2019. "Mark Zuckerberg calls for stronger regulation of internet." *The Guardian*, March 30. www.theguardian.com/technology/2019/mar/30/mark-zuckerberg-calls-for-stronger-regulation-of-internet.

Rudgard, Olivia and James Cook. 2019. "Twitter boss calls for social media regulation." *The Telegraph*, April 3. www.telegraph.co.uk/technology/2019/04/03/twitter-boss-calls-social-media-regulation/.

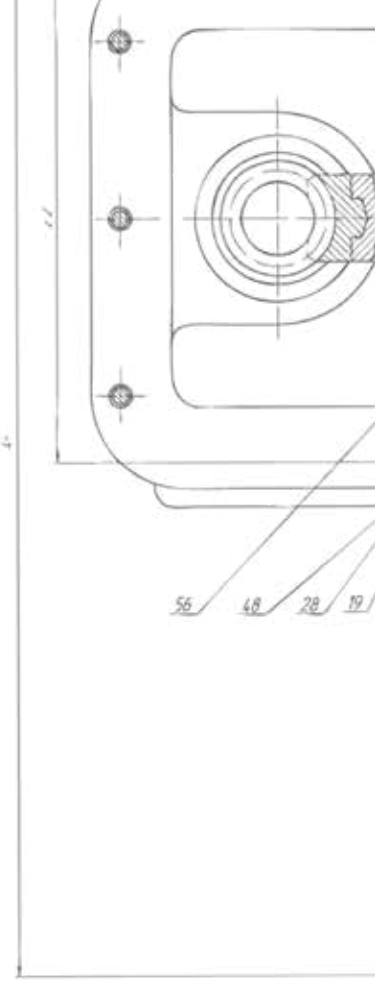
Sassòli, Marco. 2014. "Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to be Clarified." *International Law Studies* 90: 308–40.

Schmitt, Michael N. and Jeffrey S. Thurnher. 2013. "'Out of the Loop': Autonomous Weapon Systems and the Law of Armed Conflict." *Harvard National Security Journal* 4: 231–81.

Taddonio, Patrice. 2019. "How China's Government Is Using AI on Its Uighur Muslim Population." PBS.org, November 21. www.pbs.org/wgbh/frontline/article/how-chinas-government-is-using-ai-on-its-uighur-muslim-population/.

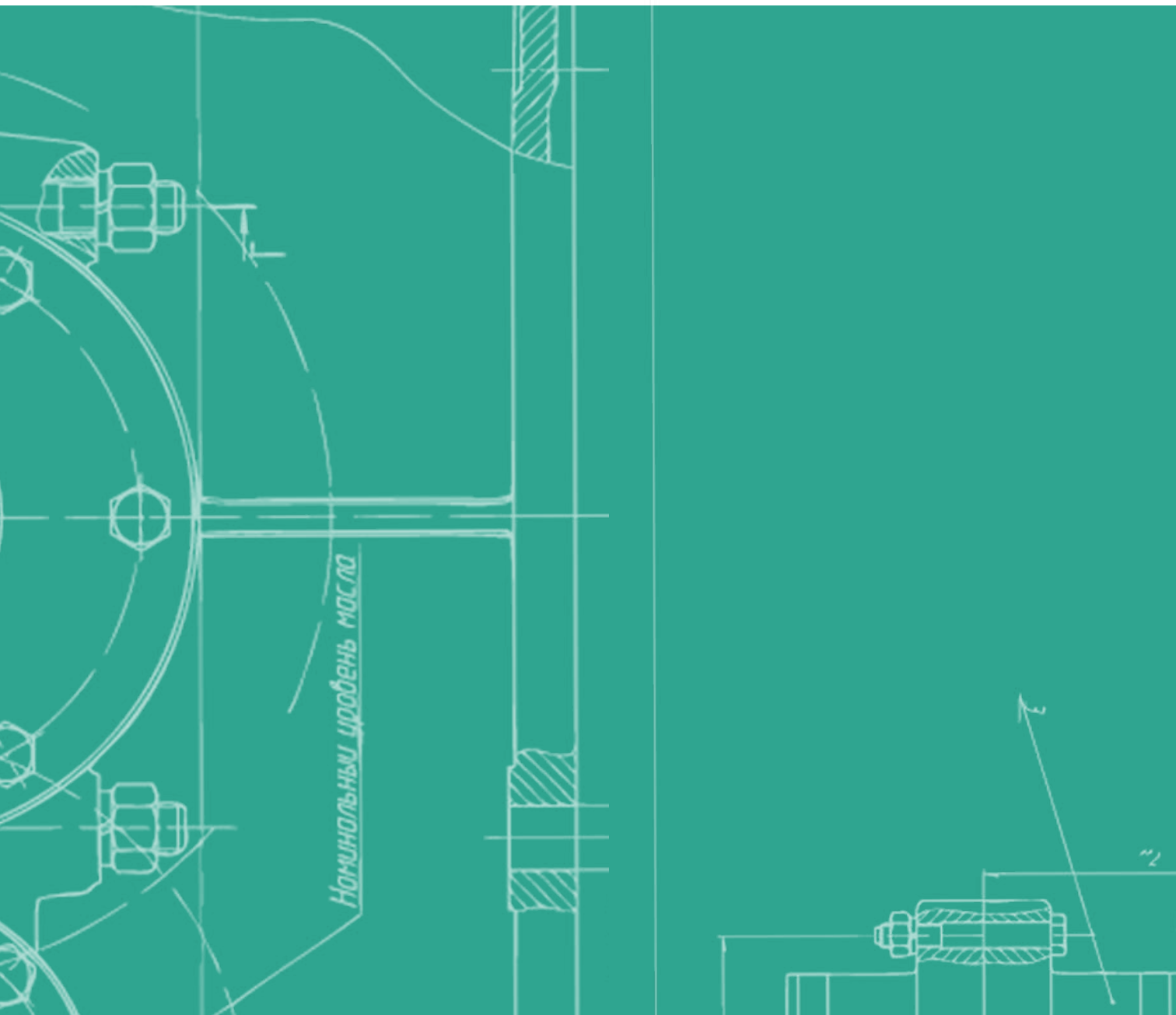
ABOUT THE AUTHOR

Liis Vihul is the founder and chief executive officer of Cyber Law International, a firm that provides international cyber law training and consulting services for governments and international organizations worldwide. Previously, she spent nine years as a senior analyst in the Law and Policy Branch at the NATO Cooperative Cyber Defence Centre of Excellence. She was a member of the Estonian delegation at the United Nations Group of Governmental Experts (GGE) on Developments in the Field of Information and Telecommunications in the Context of International Security (2014–2015 and 2016–2017) and advises the Estonian Ministry of Foreign Affairs in the current 2019–2021 UN GGE process. Liis holds a master's degree in law from the University of Tartu and a master's degree in information security from the University of London.

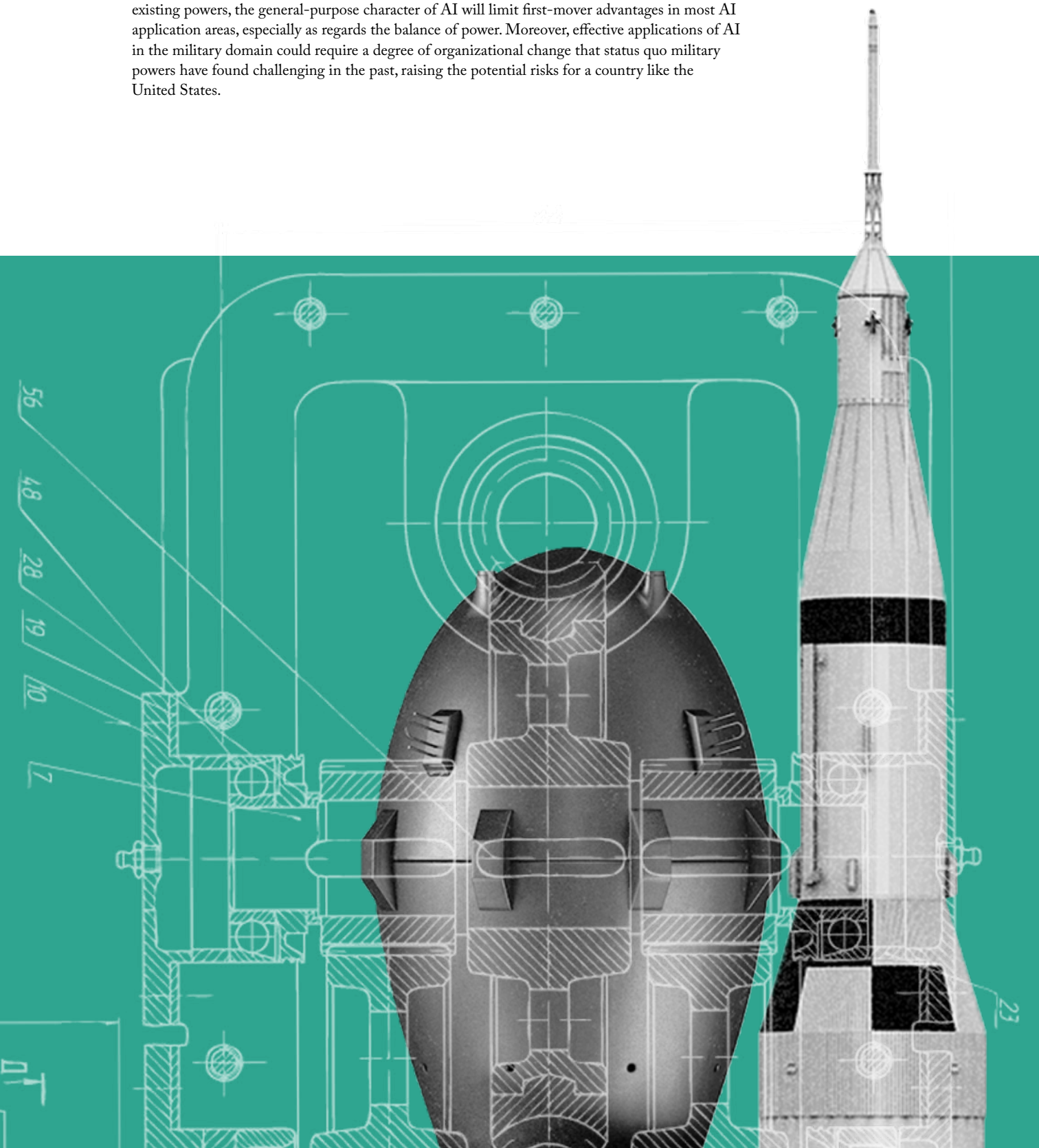


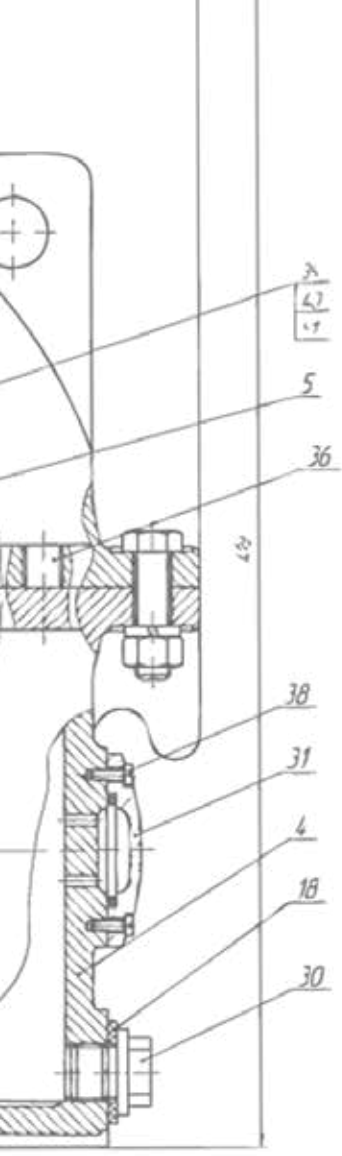
AI and the Diffusion of Global Power

Michael C. Horowitz



What role will artificial intelligence (AI) play in shaping the balance of power? AI is a general-purpose technology (GPT) with many applications across civilian and military domains. Accordingly, the impetus for AI innovation and invention also comes from a broad set of actors, with countries and companies investing heavily. The history of economic and military power suggests that while some applications of AI might enhance existing powers, the general-purpose character of AI will limit first-mover advantages in most AI application areas, especially as regards the balance of power. Moreover, effective applications of AI in the military domain could require a degree of organizational change that status quo military powers have found challenging in the past, raising the potential risks for a country like the United States.





What Is AI?

AI is not a single widget, unlike a semiconductor or even a nuclear weapon. While the specific definition is contested, AI is a universe of techniques, such as machine learning and neural networks, that involve the use of computers and computing power for tasks that we used to think required human intelligence and engagement (West 2018; Burnham 2020).

That question raises another: what is necessary to succeed in AI? The oft-used phrase that data is the new oil (Toonders 2018) is, in the context of AI, probably wrong. Building a successful algorithm requires not only having a lot of data, but also having the right data, the programming talent to write an algorithm and the computational power, or "compute," to train the algorithm (Hwang 2018). The number of cases where more data is the determining factor in predicting an algorithmic advance may be more limited than it might seem at first glance. One area where more data could matter is in predicting consumer behaviour, or, more nefariously, surveillance of a domestic population. Even then, at some point there are declining returns to gaining additional data. In a world of AI, any so-called autocratic advantage (Harari 2018) due to greater data access is likely quite limited. Nonetheless, that lack of substantive advantage over other types of regimes won't prevent autocracies from exploiting access to their own populations' data as a new tool to repress their populations more effectively.

The difference between data quantity and quality, and the importance of processing power, is critical to thinking about potential military uses of AI. On the one hand, one could argue that China has an advantage in AI because the size of its population gives it access to huge sets of population data (*The Economist* 2020a). But that data will not help China train the algorithms that are likely to be most relevant for twenty-first-century military conflicts. Instead, it's the American military's decades of experience fighting wars (whether one agrees or disagrees with the United States' involvement in those conflicts) that should yield training data pertinent to designing algorithms for logistical planning, promotion and assignments, and operations on the front lines. The potential for generative adversarial networks, or GANs, to train algorithms (Gui et al. 2020) also limits the relevance of a raw advantage in data access. GANs use simulated environments, and competition, to substitute for a lack of real-world data.

Despite the way a few companies, such as Google and Alibaba, have led in AI so far, it is unlikely that a small number of companies, or countries, will monopolize AI knowledge, particularly as AI techniques mature and become better known. Being aware that another company, or country, has designed an algorithm that can do a particular task, even without knowing how it was done, could provide vicarious knowledge that aids competitors in rapidly adopting algorithms debuted by others and make first-mover advantages relatively limited. The tight, high-end labour market in AI is likely to loosen in the coming years, particularly as universities around the world are producing a new generation of AI programmers and researchers.

Moreover, a key constraint on training algorithms, and something that could slow diffusion, is the massive computing power necessary to train cutting-edge algorithms. However, the relative cost of computing power (Hernandez and Brown 2020) is finally declining (*The Economist* 2020b), which reduces a potential barrier to mimicry.

Finally, cybersecurity will be essential for protecting algorithms from hackers and espionage. Even if hardware barriers continue to exist and countries or companies lack data to train algorithms themselves, cyberespionage could still provide a means to steal knowledge about algorithms. Through data poisoning, countries or industrial competitors could try to prevent potential adversaries from developing effective algorithms in the first place (Khurana et al. 2019). Algorithms that *have* been developed successfully are also vulnerable; through data hacking or spoofing (Heaven 2019), adversaries could prevent these effectively trained algorithms from actually being implemented (Yang et al. 2020).

AI and GPTs

GPTs are technologies with a wide number of extensive uses across many sectors (Bresnahan and Trajtenberg 1995). Historical examples include the combustion engine and electricity, while a more modern example is information technology (Jovanovic and Rousseau 2005). Coordinating innovation on GPTs is difficult,

because of the large number of actors simultaneously pursuing inventions in related, or even the same, sectors.

AI is not a new field. Symbolic approaches to algorithm development, characterized by rule-based systems known as "Good Old-Fashioned Artificial Intelligence" (Haugeland 1985), have existed for decades. But the pace of advances in AI has grown in recent years due to new approaches. AI functions as a GPT because of the number of potential sectors for its use, and the large set of actors working on algorithms (Pethokoukis 2019). Researchers around the world, both at universities and at companies, are moving forward the state of the art in the basic understanding of AI and in specific application areas. Key areas of AI include vision algorithms and text algorithms, while methods include machine learning, deep learning and neural networks (Sejnowski 2020).

AI is an especially broad technology, with potential applications encompassing everything from the algorithms that determine Netflix and Amazon recommendations to the computer vision algorithms that attempt to detect missile launches. This makes AI much more like GPTs of the past — such as the steam engine — than like a regular dual-use technology. Dual-use technologies — the Global Positioning System, for instance — can be used for either military or civilian purposes. Algorithms can also be used for either military or civilian purposes, but their breadth and diversity of potential application means the dual-use frame may be less appropriate.

If AI is a GPT, that means, on balance, its applications are likely to become diffused, rather than remain concentrated. Given that innovation in the underlying science comes from private industry and universities, rather than from classified military research (despite the key funding the Defense Advanced Research Projects Agency provided to help launch the AI field), a wide range of actors have access to information on technology breakthroughs. In contrast, stealth technology, an application area of material science, represents a classic example of a technology with purely military applications. When technologies only have military applications, the number of potentially interested actors are limited, as are the net resources available for investment. Military-only applications also make inventions more likely to diffuse slowly, due to secrecy. Research shows that technologies based on underlying commercial research, on balance, spread faster than technologies based on underlying military research (Horowitz 2010).

Given the general-purpose character of AI, and the trends described above — interest from companies around the world in AI, and declining costs in computing power — it should be relatively difficult to control the spread of capabilities built from algorithms.

AI and Organizational Change

Yet the way AI will impact the balance of power is not simply a question of how technology spreads. After all, as described above, power generally comes not from invention in and of itself, but through its uses, which require concepts of operation and organizational change to implement those visions. This is true not only when thinking about how technology can impact economic power but also when thinking about its consequences for military power. When adopting new capabilities requires doing what militaries or companies have done before, only better — like a more efficient computer — status quo actors tend to centralize and consolidate power.

If AI is a GPT, that means, on balance, its applications are likely to become diffused, rather than remain concentrated.

However, when adoption requires disruptive organizational change, it opens the potential for both significant shifts in economic power and underlying changes to the military balance of power. A classic example in military history is the aircraft carrier. When the United Kingdom's Royal Navy invented the aircraft carrier with the HMS *Furious* in 1918, it viewed the utility of the aircraft carrier primarily as

an aerial spotter for the battleship. Because the Royal Navy was the best in the world at battleship warfare, it thought about aircraft carriers as a way to improve an already well-established competency. Alternatively, the United States Navy and the Japanese Navy, in part due to their need to project power across the vast Pacific Ocean, thought about the aircraft carrier more as a mobile airfield. The United States, in particular, reorganized its navy in World War II to take advantage of the striking power of naval aircraft launched from aircraft carriers, transforming naval warfare as a result. The Royal Navy, in contrast, bound to battleships due to organizational politics and the weight of history, fell behind.

The United States, as the leading military in the world, is both a role model and a target.

Given that AI is a GPT with many areas of use, different applications of AI may require different types of organizational change to take advantage of them. For example, a shift by air forces from focusing on low numbers of capital-intensive aircraft, such as the F-35 fighters, with highly trained pilots on board, to low-cost drone swarms — operating as a pack and uninhabited, with one pilot overseeing many aircraft — would be extremely disruptive, organizationally, for a military such as the United States'. In contrast, using computer vision algorithms to better identify patterns and detect missile launches or assist humans in identifying targets would not be as disruptive. But it is also important to keep in mind that most uses of AI by militaries will not be on the battlefield. Instead, they will be in logistics, personnel and other arenas far from the fight, but still potentially very consequential to overall military effectiveness.

AI and the Balance of Power

The large degree of uncertainty surrounding applications of AI by militaries makes determining the impact of AI on the balance of power difficult. However, some possibilities can be forecast, given the diverse potential military uses of algorithms and some of the general tendencies of the AI field.

Imagine two different types of military uses of AI. The first, and most common, use of AI by militaries will be general-purpose applications based on related algorithms in the commercial world. Project Maven in the United States (Seligman 2018), which draws on computer vision algorithms developed by companies for non-military purposes, exemplifies one general-purpose-derived application of AI by militaries. Military applications will require more cybersecurity, and some specialization, but the underlying basis of the algorithms will be similar. Thus, in these application areas, first-mover advantages should be relatively limited. Countries with substantial militaries and information economies should be able to mimic advances relatively quickly, since the underlying technology will be relatively accessible. Thus, these uses of AI should not, on their own, have a large relative impact on the balance of power. However, even if the technology is mimicked relatively quickly, the impact on the balance of power could still be asymmetric, as bureaucratic politics mean some militaries are better poised than others to take advantage.

More specialized applications of AI for militaries, although less frequent, could create much larger first-mover advantages and have important consequences for the balance of power. Algorithms designed to help human commanders manage a complex and multi-dimensional battlespace, for example, do not have as many obvious commercial corollaries. Thus, militaries are more likely to invest in the science required for breakthroughs, and that research will likely be secret and harder to copy by potential adversaries (although there would still be the potential for mimicry after seeing algorithms that others debut).

The United States, as the leading military in the world, is both a role model and a target. There is a great deal of rhetoric in the United States surrounding investments in AI, but despite the creation of the Joint Artificial Intelligence Center, there is concern that the rhetoric is not matched by the budgetary reality of limited investments. Moreover, as the leading military power in the world, the United States, like the Royal Navy with aircraft carriers, arguably faces the biggest risk. Meanwhile, even though China's aspirations

to leverage AI to leapfrog the United States' economy, and the American military, are clear, it is much less clear whether Chinese investments will translate into surpassing the United States in AI, let alone with applications relevant for the balance of power. Moreover, around the world, from Canada to Israel to Singapore, governments are ramping up their AI investments and considering potential military uses. As the pandemic of coronavirus disease 2019 continues, one potential consequence of workplaces being unsafe for humans may be to accelerate investments in robotics and autonomous systems. This possibility could apply to the military and the private sector, although the consequences to the civilian economy will likely be clearer first.

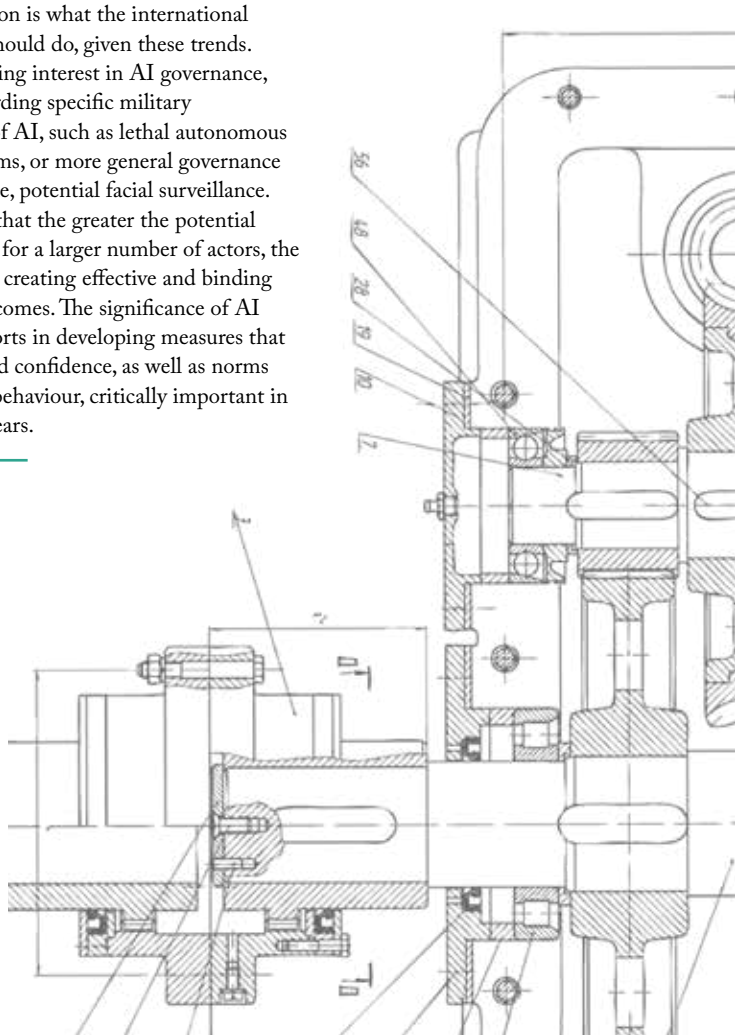
Finally, this evaluation of the way AI could shape the balance of power, and the extent to which it might concentrate or diffuse power, focuses on so-called narrow applications of AI. Narrow algorithms are built to do one thing, such as play a game; an example is AlphaGo Zero, software developed in 2017 by DeepMind to play Go and trained with reinforcement learning, meaning that it learned to play the game without being fed training data from human game play. The impact of AI on the balance of power could be different if one company or country achieves a massive breakthrough that enables the creation of artificial general intelligence. A general algorithm that could write other algorithms, operate in many domains and avoid the problem of catastrophic forgetting (forgetting previous learning after acquiring new information in a different area) would give a first mover a substantial advantage. Some, such as Nick Bostrom (2014), director of the Future of Humanity Institute at Oxford University, worry that the first-mover advantages might be so large that they would be calamitous. Thus, the consequences on the balance of power would be very different.

Conclusion

The tremendous uncertainty among experts surrounding the potential for advances in AI (Grace et al. 2018) makes forecasting the consequences on the balance of power difficult. Nevertheless, investments by militaries around the world, and concern on the part of many researchers and organizations interested in understanding potential changes in the conduct of warfare, mean it is important to understand the likely impact of AI now. If AI is like other GPTs, it will certainly create winners and losers based on the ability and capacity of countries and companies to effectively use AI, in particular on their ability to secure algorithms from data poisoning, hacking and spoofing, which will reduce the risk of accidents.

But GPTs, as technology categories that are broader than specific dual-use widgets, tend to diffuse relatively quickly, especially in comparison to purely military technologies. In an absolute sense, algorithms, and knowledge of how to design them, are also likely to diffuse relatively quickly (compared to, say, knowledge about how to build an F-35). A big question, though, is the extent to which taking advantage of AI, whether more general or more specialized applications, will require significant, disruptive, organizational change. The higher the degree of change required, history suggests, the greater the potential for a shift in the balance of power (Horowitz 2010), and the greater the risk for a leading military such as the United States.

A final question is what the international community should do, given these trends. There is growing interest in AI governance, whether regarding specific military applications of AI, such as lethal autonomous weapon systems, or more general governance of, for example, potential facial surveillance. A paradox is that the greater the potential impact of AI, for a larger number of actors, the more difficult creating effective and binding regulation becomes. The significance of AI will make efforts in developing measures that build trust and confidence, as well as norms surrounding behaviour, critically important in the coming years.



WORKS CITED

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford, UK: Oxford University Press.

Bresnahan, Timothy F. and M. Trajtenberg. 1995. "General purpose technologies: 'Engines of growth'?" *Journal of Econometrics* 65 (1): 83–108.

Burnham, Kristin. 2020. "Artificial Intelligence vs. Machine Learning: What's the Difference?" *Northeastern University* (blog), May 6. www.northeastern.edu/graduate/blog/artificial-intelligence-vs-machine-learning-whats-the-difference/.

Grace, Katja, John Salvatier, Allan Dafoe, Baobao Zhang and Owain Evans. 2018. "Viewpoint: When Will AI Exceed Human Performance? Evidence from AI Experts." *Journal of Artificial Intelligence Research* 62: 729–54. https://doi.org/10.1613/jair.1.11222.

Gui, Jie, Zhenan Sun, Yonggang Wen, Dacheng Tao and Jieping Ye. 2020. "A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications." Cornell University arXiv e-print, January 20. https://arxiv.org/abs/2001.06937.

Harari, Yuval Noah. 2018. "Why Technology Favors Tyranny." *The Atlantic*, October. www.theatlantic.com/magazine/archive/2018/10/yuval-noah-harari-technology-tyranny/568330/.

Haugeland, John. 1985. *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press.

Heaven, Douglas. 2019. "Why deep-learning AIs are so easy to fool." *Nature*, October 9. www.nature.com/articles/d41586-019-03013-5.

Hernandez, Danny and Tom B. Brown. 2020. "Measuring the Algorithmic Efficiency of Neural Networks." Cornell University arXiv e-print, May 8. https://arxiv.org/abs/2005.04305.

Horowitz, Michael C. 2010. *The Diffusion of Military Power: Causes and Consequences for International Politics*. Princeton, NJ: Princeton University Press.

Hwang, Tim. 2018. "Computational Power and the Social Impact of Artificial Intelligence." March 23. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3147971.

Jovanovic, Boyan and Peter L. Rousseau. 2005. "General Purpose Technologies." In *Handbook of Economic Growth, Volume 1B*, edited by Philippe Aghion and Steven N. Durlauf, 1181–1224. Amsterdam, The Netherlands: North Holland.

Khurana, N., S. Mittal, A. Piplai and A. Joshi. 2019. "Preventing Poisoning Attacks On AI Based Threat Intelligence Systems." 2019 IEEE 29th International Workshop on Machine Learning for Signal Processing, Pittsburgh, PA, October 13–16. https://ieeexplore.ieee.org/document/891880.

Pethokoukis, James. 2019. "How AI is like that other general purpose technology, electricity." *AEIdeas* (blog), November 25. Washington, DC: American Enterprise Institute. www.aei.org/economics/how-ai-is-like-that-other-general-purpose-technology-electricity/.

Sejnowski, Terrence J. 2020. "The unreasonable effectiveness of deep learning in artificial intelligence." *Proceedings of the National Academy of Sciences of the United States of America*, January 28. www.pnas.org/content/early/2020/01/23/1907373117.

Seligman, Lara. 2018. "Pentagon's AI Surge on Track, Despite Google Protest." *Foreign Policy*, June 29. https://foreignpolicy.com/2018/06/29/google-protest-wont-stop-pentagons-a-i-revolution/.

The Economist. 2020a. "China's success at AI has relied on good data." January 2. www.economist.com/technology-quarterly/2020/01/02/chinas-success-at-ai-has-relied-on-good-data.

———. 2020b. "The cost of training machines is becoming a problem." June 11. www.economist.com/technology-quarterly/2020/06/11/the-cost-of-training-machines-is-becoming-a-problem.

Toonders, Joris. 2018. "Data Is the New Oil of the Digital Economy." *Wired*. www.wired.com/insights/2014/07/data-new-oil-digital-economy/.

West, Darrell M. 2018. "What is artificial intelligence?" Brookings Institution, October 4. www.brookings.edu/research/what-is-artificial-intelligence/.

Yang, Chao-Han Huck, Jun Qi, Pin-Yu Chen, Yi Ouyang, I-Te Danny Hung, Chin-Hui Lee and Xiaoli Ma. 2020. "Enhanced Adversarial Strategically-Timed Attacks Against Deep Reinforcement Learning." Paper presented at the International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, May 4–8. https://ieeexplore.ieee.org/document/9053342.

ABOUT THE AUTHOR

Michael C. Horowitz is Richard Perry Professor and the director of Perry World House at the University of Pennsylvania. Michael is the author of *The Diffusion of Military Power: Causes and Consequences for International Politics*, and the co-author of *Why Leaders Fight*. He won the 2017 Karl Deutsch Award given by the International Studies Association for early career contributions to the fields of international relations and peace research. His research interests include the intersection of emerging technologies such as AI and robotics with global politics, military innovation, the role of leaders in international politics and geopolitical forecasting methodology. Michael previously worked for the Office of the Under Secretary of Defense for Policy in the US Department of Defense. He is affiliated with the Center for a New American Security, the Center for Strategic and International Studies, and the Foreign Policy Research Institute. He is a member of the Council on Foreign Relations. He received his Ph.D. in government from Harvard University and his B.A. in political science from Emory University. You can find him on Twitter @mchorowitz.

Governing Cyberspace during a Crisis in Trust

A CIGI essay series on the economic potential – and vulnerability – of transformative technologies and cyber security

While technology has led to convenience, efficiency and wealth creation, the push to digitize society quickly and relentlessly has left the core of the global economic model vulnerable.

cigionline.org/cyberspace


Номинальный уровень масла

Influence Operations and Disinformation on Social Media

Samantha Bradshaw

Amid the coronavirus disease 2019 (COVID-19) pandemic, foreign state actors have been spreading disinformation on social media about the disease and the virus that causes it (Bright et al. 2020; Molter 2020). Covering a variety of topics — from its origin to potential cures, or its impact on Western societies — the creation and dissemination of COVID-19 disinformation online has become widespread.

States — such as Russia and China — have taken to Facebook, Twitter and YouTube to create and amplify conspiratorial content designed to undermine trust in health officials and government administrators, which could ultimately worsen the impact of the virus in Western societies (Barnes and Sanger 2020).

Although COVID-19 has highlighted new and incredible challenges for our globalized society, foreign influence operations that capitalize on moments of global uncertainty are far from new. Over the past few years, public and policy attention has focused largely on foreign influence operations targeting elections and referendums, but health-related conspiracy theories created and amplified as part of state propaganda campaigns also have a long history.

One example is the conspiracy theory that AIDS (acquired immune deficiency syndrome) was the result of a biological weapons experiment conducted by the US government. Historians have documented how Soviet operatives leaked "evidence" into foreign institutions and media outlets questioning the origin of the virus (Boghardt 2009). Because the US government was slow to respond to the AIDS epidemic, which disproportionately affected gay men and people of colour, conspiracy theories about its origin heightened suspicions within these communities that the US government was responsible (Qiu 2017). Decades later, public health research has shown that many people still hold conspiratorial beliefs about the human immunodeficiency virus (HIV) that causes AIDS, which has negatively affected treatment for the disease (Bogart et al. 2010).

Social media platforms have come to dominate almost every aspect of human interaction, from interpersonal relations to the global economy.

Part of the reason why the HIV/AIDS conspiracy was effectively inculcated into the belief systems of everyday people was because it involved identifying and exploiting pre-existing divisions among society and then using disinformation to sow further discord and distrust. Today, state actors have applied the same playbook used during the Cold War as part of contemporary foreign influence operations: in the lead-up to the 2016 US presidential election, for example, disinformation and conspiracy theories injected into social and mainstream media were used to exacerbate racial tensions in the United States, particularly around the Black Lives Matter movement (DiResta et al. 2018; Howard et al. 2018), but also around religious (Hindman and Barash 2018) and gender divides (Bradshaw 2019).

What has changed from the Cold War–era information warfare to contemporary influence operations is the information and media landscape through which disinformation can be circulated. Innovations in technology have transformed modern-day conflict and the ways in which foreign influence operations take place. Over the past two decades, state and non-state actors have increasingly used the internet to pursue political and military agendas, by combining traditional military operations with cyberattacks and online propaganda campaigns (North Atlantic Treaty Organization 2016). These "hybrid methods" often make use of the spread of disinformation to erode the truth and undermine the credibility of international institutions and the liberal world order (National Defence Canada 2017).

Today, unlike in the past, when disinformation campaigns were slow, expensive and data-poor, social media provides a plethora of actors with a quick, cheap and data-rich medium to use to inject disinformation into civic conversations. Algorithms that select, curate and control our information environment might prioritize information based on its potential for virality, rather than its grounding in veracity. Behind the veil of anonymity, state-sponsored trolls can bully, harass and prey on individuals or communities online, discouraging the expression of some of the most important voices in activism and journalism. Sometimes the people behind these accounts are not even real, but automated scripts of code designed to amplify propaganda, conspiracy and disinformation online. The very design of social media technologies can enhance the speed, scale and reach of propaganda and disinformation, engendering new international security concerns around foreign influence operations online.

Foreign Influence Operations in a Platform Society

From public health conspiracies to disinformation about politics, social media has increasingly become a medium used by states to meddle in the affairs of others (Bradshaw and Howard 2018; 2019). From China's disinformation campaigns that painted Hong Kong democracy protestors as violent and unpopular dissidents (Wong, Shepherd and Liu 2019), to Iranian-backed disinformation campaigns targeting political rivals in the Gulf (Elswah, Howard and Narayanan 2019), state actors are turning to social media as a tool of geopolitical influence. And it is not just state actors who turn to social media platforms to spread disinformation and propaganda. Populist political parties, far-right media influencers, dubious strategic communications firms and the charlatans of scientific disinformation have all found a home for conspiracy, hate and fear on social media (Campbell-Smith and Bradshaw 2019; Evangelista and Bruno 2019; Numerato et al. 2019). What is it about the contemporary communication landscape that makes social media such a popular — and arguably powerful — platform for disinformation?

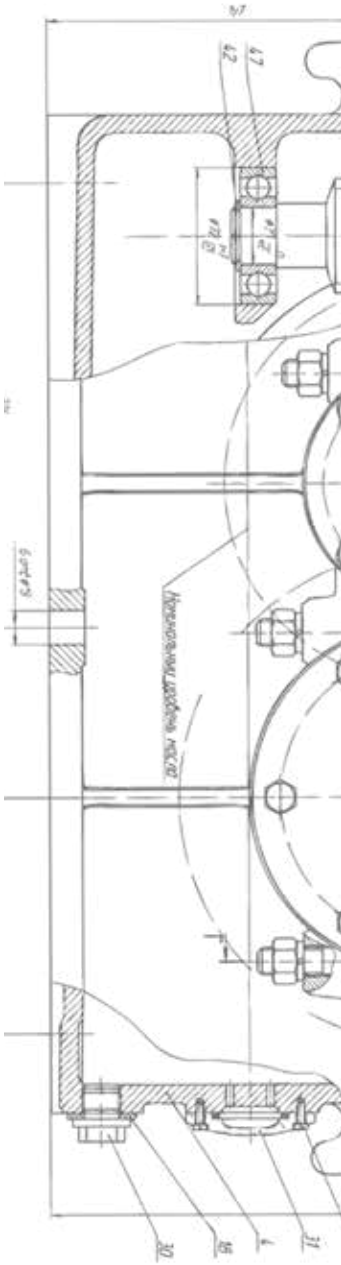
Social media platforms have come to dominate almost every aspect of human interaction, from interpersonal relations to the global economy. But they also perform important civic functions. Increasingly, these platforms are an important source of news and information for citizens around the world (Newman et al. 2020). They are a place for political discussion and debate, and for mobilizing political action (Benkler 2007; Castells 2007; Conover et al. 2013). Politicians also rely on social media for political campaigning, galvanizing support and connecting with their constituents (Hemsley 2019; Howard 2006; Kreiss 2016). But social media platforms are not neutral platforms (Gillespie 2010). Scholars have described how their technical designs and governance policies (such as terms of service, community standards or advertising policies) embed a wide range of public policy concerns, from freedom of speech and censorship to intellectual property rights and fair use or tensions between privacy and surveillance online (DeNardis and Hackl 2015; Gillespie 2019; Hussain and Howard 2013; MacKinnon 2012). Platform design and governance also impact the democratic functions of platforms, including how disinformation and propaganda are spread. While it is important to recognize that all technologies have socio-political implications to various degrees, several characteristics of social media platforms create a particular set of concerns for the spread of disinformation and propaganda.

AGGREGATION

One of the most salient features of today's information and communication environment is the massive amount of data aggregated about individuals and their social behaviour. The immense amount of data we leave behind as we interact with technology and content has been called "data exhaust" by some scholars (Deibert 2015). Our exhaust — or the by-product of our interactions with online content — is used by platforms to create detailed pictures of who we are not only as people and consumers, but also as citizens or potential voters in a democracy (Tufekci 2014). The collection, aggregation and use of data allows foreign adversaries to micro-target users with political advertisements during elections. Like all political advertising, these messages could drive support and mobilization for a certain candidate or suppress the political participation of certain segments of the population (Baldwin-Philippi 2017; Chester and Montgomery 2019; Kreiss 2017). We have already seen foreign agents purchase political advertisements to target individuals or communities with messages of mobilization and suppression (Mueller 2019). Although platforms have taken several steps to limit foreign advertising on their platforms, such as currency restrictions or account verification measures, foreign actors have found ways to subvert these measures (Satariano 2018).

ALGORITHMS

Platforms apply algorithms — or automated sets of rules or instructions — to transform data into a desired output. Using mathematical formulas, algorithms rate, rank, order and deliver content based on factors such as an individual user's data and personal preferences (Bennett 2012), aggregate trends in the interests and behaviour of similar users (Nahon and Hemsley 2013), and reputation systems that evaluate the quality of information (van Dijck, Poell and de Waal 2018). The algorithmic curation of content — whether it be a result of personalization, virality



With anonymity, there is a lack of transparency about the source of information and whether news, comments or debate come from authentic voices or ones trying to distort the public sphere.

and trends, or reputation scores — affects how news and information is prioritized and delivered to users, including whether algorithms present diverse views or reinforce singular ones (Borgesius et al. 2016; Dubois and Blank 2018; Flaxman, Goel and Rao 2016; Fletcher and Nielsen 2017), nudge users toward extreme or polarizing information (Horta Ribeiro et al. 2019; Tufekci 2018) or emphasize sensational, tabloid or junk content over news and other authoritative sources of information (Bradshaw et al. 2020; Neudert, Howard and Kollanyi 2019).

ANONYMITY

Platforms afford different levels of anonymity to users. Whether users must use their real name has implications for whether bots, trolls or even foreign state actors use anonymity to mask their identity in order to harass or threaten political activists and journalists, or to distort authentic conversations about politics (Nyst and Monaco 2018). With anonymity, there is a lack of transparency about the source of information and whether news, comments or debate come from authentic voices or ones trying to distort the public sphere. Related to the question of anonymity is the question of data disclosure and how personal data disclosed to third parties can be used if unscrupulous firms or foreign state actors are able to use psychographic profiles to suppress voter turnout (Wylie 2020).

AUTOMATION

Platforms afford automation — where accounts can automatically post, share or engage with content or users online. Unlike a human user, automated accounts — which are sometimes referred to as "political bots" or "amplifier accounts" — can post much more frequently and consistently than any human user (McKelvey and Dubois 2017). Although there are many ways to classify automated accounts and the activities they perform (Gorwa and Guilbeault 2020), they generally perform two functions when it comes to foreign influence operations. First, by liking, sharing, retweeting or posting content, automated accounts can generate a false sense of popularity, momentum or relevance around a particular person or idea. Networks of bots can be used to distort conversations online by getting disinformation or propaganda to trend (Woolley 2016). Second, automation has been an incredibly powerful tool in the targeting and harassment of journalists and activists, whereby individuals are flooded with threats and hate by accounts that are not even real (Nyst and Monaco 2018).

The Future of Disinformation and Foreign Influence Operations

In conclusion, the spread of disinformation and propaganda online are growing concerns for the future of international security. The salient features of platforms — aggregation, algorithms, anonymity and automation — are some of the ways contemporary technologies can contribute to the spread of harmful content online, and foreign state actors are increasingly leveraging these tools to distort the online public sphere. The use of social media for "hybrid" methods of warfare is a broader reflection on how technological innovation changes the nature of conflict. Indeed, technology has always been recognized as a force that enables social and political transformation (Nye 2010). Similarly, the unique features of our contemporary information and communication environment provide new opportunities for state actors to use non-traditional methods of warfare to pursue their goals.

As we see innovations in technology, we will also see innovations in the way in which propaganda and disinformation spread online. The Internet of Things, which is already

revolutionizing the way we live, creates even more data about us as individuals and as citizens. What happens in a world where we can measure someone's physiological response to propaganda through wearable technology? We interact with "chatbots" like Alexa and Siri every day. What happens when the growing sophistication of chatbot technology is applied to political bots on Facebook or Twitter? How will the platforms differentiate between genuine human conversations and automated interactions?

Thus far, combatting disinformation and propaganda has been a constant game of whack-a-mole. Private responses focus on third-party fact-checking or labelling information that might be untrustworthy, misleading or outright false. In the form of laws and regulations, governments place a greater burden on platforms to remove certain kinds of harmful content, often without defining what constitutes harm. But propaganda and disinformation are also *systems problems*. Too often, public and private responses focus on the content. However, these responses ignore the technical agency of platforms to shape, curate and moderate our information ecosystem. Rather than focusing solely on the content, we need to look at the deeper systemic issues that make disinformation and propaganda go viral in the first place. This means thinking about the features of platforms that enhance or exacerbate the spread of harmful content online.

WORKS CITED

Baldwin-Philippi, J. 2017. "The Myths of Data-Driven Campaigning." *Political Communication* 34 (4): 627–33. https://doi.org/10.1080/10584609.2017.1372999.

Barnes, J. E. and D. E. Sanger. 2020. "Russian Intelligence Agencies Push Disinformation on Pandemic." *The New York Times*, July 28. www.nytimes.com/2020/07/28/us/politics/russia-disinformation-coronavirus.html.

Benkler, Y. 2007. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. New Haven, CT: Yale University Press.

Bennett, W. L. 2012. "The Personalization of Politics: Political Identity, Social Media, and Changing Patterns of Participation." *The ANNALS of the American Academy of Political and Social Science* 644 (1): 20–39. https://doi.org/10.1177/0002716212451428.

Bogart, L. M., G. Wagner, F. H. Galvan and D. Banks. 2010. "Conspiracy Beliefs About HIV Are Related to Antiretroviral Treatment Nonadherence Among African American Men With HIV." *Journal of Acquired Immune Deficiency Syndrome* 53 (5): 648–55. https://doi.org/10.1097/QAI.0b013e3181c57dbc.

Boghardt, T. 2009. "Operation INFEKTION: Soviet Bloc Intelligence and Its AIDS Disinformation Campaign." *Studies in Intelligence* 53 (4): 1–24.

Borgesius, F. J. Z., D. Trilling, J. Möller, B. Bodó, C. H. de Vreese and N. Helberger. 2016. "Should we worry about filter bubbles?" *Internet Policy Review* 5 (1): 1–16. https://policyreview.info/node/401/pdf.

Bradshaw, S. 2019. *The Gender Dimensions of Foreign Influence Operations*. Report prepared at the request of Global Affairs Canada.

Bradshaw, S. and P. N. Howard. 2018. "Challenging Truth and Trust: A Global Inventory of Organized Social Media Manipulation." Working Paper 2018.1. Oxford, UK: Project on Computational Propaganda, Oxford Internet Institute.

———. 2019. "The Global Disinformation Order: 2019 Global Inventory of Organised Social Media Manipulation." Working Paper 2019.2. Oxford, UK: Project on Computational Propaganda, Oxford Internet Institute. https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/09/CyberTroop-Report19.pdf.

Bradshaw, S., P. N. Howard, B. Kollanyi and L.-M. Neudert. 2020. "Sourcing and Automation of Political News and Information over Social Media in the United States, 2016–2018." *Political Communication* 37 (2): 173–93.

Bright, J., H. Au, H. Bailey, M. Elswah, M. Schliebs, N. Marchal, C. Schwieter, K. Rebello and P. N. Howard. 2020. "Coronavirus Coverage by State-Backed English-Language News Sources: Understanding Chinese, Iranian, Russian and Turkish Government Media." Data Memo 2020.2. Oxford, UK: Project on Computational Propaganda, Oxford Internet Institute. https://comprop.oii.ox.ac.uk/research/posts/coronavirus-coverage-by-state-backed-english-language-news-sources/.

Campbell-Smith, U. and S. Bradshaw. 2019. "Global Cyber Troops Country Profile: India." Oxford, UK: Project on Computational Propaganda, Oxford Internet Institute. https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/05/India-Profile.pdf.

Castells, M. 2007. "Communication, Power and Counter-power in the Network Society." *International Journal of Communication* 1 (1): 29. https://ijoc.org/index.php/ijoc/article/view/46.

Chester, J. and K. C. Montgomery. 2019. "The digital commercialisation of US politics — 2020 and beyond." *Internet Policy Review* 8 (4). https://policyreview.info/articles/analysis/digital-commercialisation-us-politics-2020-and-beyond.

Conover, M. D., E. Ferrara, F. Menczer and A. Flammini. 2013. "The Digital Evolution of Occupy Wall Street." *PLOS ONE* 8 (5): e64679. https://doi.org/10.1371/journal.pone.0064679.

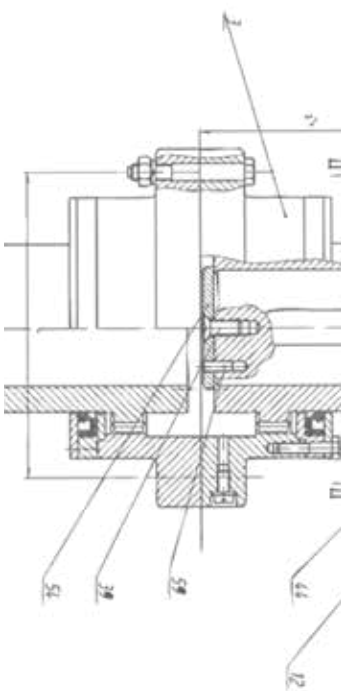
Deibert, R. 2015. "The Geopolitics of Cyberspace After Snowden." *Current History* 114 (768): 9–15. https://doi.org/10.1525/curh.2015.114.768.9.

DeNardis, L. and A. M. Hackl. 2015. "Internet Governance *by* Social Media Platforms." *Telecommunications Policy* 39 (9): 761–70. https://doi.org/10.1016/j.telpol.2015.04.003.

DiResta, R., K. P. Shaffer, B. Ruppel, D. M. Sullivan, R. Matney, R. Fox, J. Albright, B. E. Johnson. 2018. "The Tactics & Tropes of the Internet Research Agency." White paper. Austin, TX: New Knowledge. https://disinformationreport.blob.core.windows.net/disinformation-report/NewKnowledge-Disinformation-Report-Whitepaper.pdf.

Dubois, E. and G. Blank. 2018. "The echo chamber is overstated: the moderating effect of political interest and diverse media." *Information, Communication & Society* 21 (5): 729–45. https://doi.org/10.1080/1369118X.2018.1428656.

Elswah, M., P. N. Howard and V. Narayanan. 2019. "Iranian Digital Interference in the Arab World." Data Memo 2019.1. Oxford, UK: Project on Computational Propaganda, Oxford Internet Institute. https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/04/Iran-Memo.pdf.



Evangelista, R. and F. Bruno. 2019. "WhatsApp and political instability in Brazil: targeted messages and political radicalization." *Internet Policy Review* 8 (4). https://policyreview.info/articles/analysis/whatsapp-and-political-instability-brazil-targeted-messages-and-political.

Flaxman, S., S. Goel and J. M. Rao. 2016. "Filter Bubbles, Echo Chambers, and Online News Consumption." *Public Opinion Quarterly* 80 (S1): 298–320. https://doi.org/10.1093/poq/nfw006.

Fletcher, R. and R. K. Nielsen. 2017. "Are people incidentally exposed to news on social media? A comparative analysis." *New Media & Society* 20 (7): 2450–68. https://doi.org/10.1177/1461444817724170.

Gillespie, T. 2010. "The politics of 'platforms.'" *New Media & Society* 12 (3): 347–64.

———. 2019. *Custodians of the Internet: Platforms, Content Moderation and the Hidden Decisions that Shape Social Media*. New Haven, CT: Yale University Press.

Gorwa, R. and D. Guilbeault. 2020. "Unpacking the Social Media Bot: A Typology to Guide Research and Policy." *Policy & Internet* 12 (2): 225–48. https://doi.org/10.1002/poi3.184.

Hemsley, J. 2019. "Followers Retweet! The Influence of Middle-Level Gatekeepers on the Spread of Political Information on Twitter." *Policy & Internet* 11 (3): 280–304. https://doi.org/10.1002/poi3.202.

Hindman, M. and V. Barash. 2018. *Disinformation, 'Fake News' and Influence Campaigns on Twitter*. Miami, FL: Knight Foundation. https://knightfoundation.org/reports/disinformation-fake-news-and-influence-campaigns-on-twitter.

Horta Ribeiro, M., R. Ottoni, R. West, V. A. F. Almeida and W. Meira. 2019. "Auditing Radicalization Pathways on YouTube." Cornell University arXiv e-print, December 4. https://arxiv.org/abs/1908.08313v3.

Howard, P. N. 2006. *New Media Campaigns and the Managed Citizen*. Cambridge, UK: Cambridge University Press.

Howard, P. N., B. Ganesh, D. Liotsiou, J. Kelly and C. François. 2018. *The IRA, Social Media and Political Polarization in the United States*, 2012–2018. Oxford, UK: Project on Computational Propaganda, Oxford Internet Institute. https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2018/12/The-IRA-Social-Media-and-Political-Polarization.pdf.

Hussain, M. M. and P. N. Howard, eds. 2013. *State Power 2.0: Authoritarian Entrenchment and Political Engagement Worldwide*. Farnham, UK: Ashgate Publishing.

Kreiss, D. 2016. *Prototype Politics: Technology-Intensive Campaigning and the Data of Democracy*. New York, NY: Oxford University Press.

———. 2017. "Micro-targeting, the quantified persuasion." *Internet Policy Review* 6 (4). https://policyreview.info/articles/analysis/micro-targeting-quantified-persuasion.

MacKinnon, R. 2012. *Consent of the Networked: The Worldwide Struggle for Internet Freedom*. New York, NY: Basic Books.

McKelvey, F. and E. Dubois. 2017. "Computational Propaganda in Canada: The Use of Political Bots." Working Paper 2017.6. Oxford, UK: Project on Computational Propaganda, Oxford Internet Institute. http://comprop.oii.ox.ac.uk/wp-content/uploads/sites/89/2017/06/Comprop-Canada.pdf.

Molter, V. 2020. "Virality Project (China): Pandemics & Propaganda." Cyber News, March 19. Stanford, CA: Cyber Policy Center. https://cyber.fsi.stanford.edu/news/chinese-state-media-shapes-coronavirus-convo.

Mueller, R. S. I. 2019. *Report On The Investigation Into Russian Interference In The 2016 Presidential Election. Volume I of II*. Washington, DC: US Department of Justice. www.justice.gov/storage/report.pdf.

Nahon, K. and J. Hemsley. 2013. *Going Viral*. Cambridge, UK: Polity Press.

National Defence Canada. 2017. *Strong, Secure, Engaged: Canada's Defence Policy*. Ottawa, ON: National Defence. http://dgpaapp.forces.gc.ca/en/canada-defence-policy/docs/canada-defence-policy-report.pdf.

Neudert, L.-M., P. Howard and B. Kollanyi. 2019. "Sourcing and Automation of Political News and Information During Three European Elections." *Social Media + Society* 5 (3). https://doi.org/10.1177/2056305119863147.

Newman, N., R. Fletcher, A. Schulz, S. Andi and R. Kleis Nielsen. 2020. *Reuters Institute Digital News Report*. Oxford, UK: Reuters Institute for the Study of Journalism.

North Atlantic Treaty Organization. 2016. *Social Media as a Tool of Hybrid Warfare*. Riga, Latvia: North Atlantic Treaty Organization. www.stratcomcoe.org/social-media-tool-hybrid-warfare.

Numerato, D., L. Vochocová, V. Štětka and A. Macková. 2019. "The vaccination debate in the 'post-truth' era: social media as sites of multi-layered reflexivity." *Sociology of Health & Illness* 41 (S1): 82–97. https://doi.org/10.1111/1467-9566.12873.

Nye, J. S. 2010. "Cyber Power." Paper, May. Cambridge, MA: Belfer Center for Science and International Affairs, Harvard Kennedy School.

Nyst, C. and N. Monaco. 2018. *State-Sponsored Trolling: How Governments Are Deploying Disinformation as Part of Broader Digital Harassment Campaigns*. Palo Alto, CA: Institute for the Future.

Qiu, L. 2017. "Fingerprints of Russian Disinformation: From AIDS to Fake News." *The New York Times*, December 12. www.nytimes.com/2017/12/12/us/politics/russian-disinformation-aids-fake-news.html.

Satariano, A. 2018. "Ireland's Abortion Referendum Becomes a Test for Facebook and Google." *The New York Times*, May 25. www.nytimes.com/2018/05/25/technology/ireland-abortion-vote-facebook-google.html.

Tufekci, Z. 2014. "Engineering the public: Big data, surveillance and computational politics." *First Monday* 19 (7). https://doi.org/10.5210/fm.v19i7.4901.

———. 2018. "YouTube, the Great Radicalizer." *The New York Times*, March 10. www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html.

van Dijck, J., T. Poell and M. de Waal. 2018. *The Platform Society: Public Values in a Connective World*. New York, NY: Oxford University Press.

Wong, S.-L., C. Shepherd and Q. Liu. 2019. "Old messages, new memes: Beijing's propaganda playbook on the Hong Kong protests." *Financial Times*, September 3. www.ft.com/content/7ed90e60-ce89-11e9-99a4-b5ded7a7fe3f.

Woolley, S. C. 2016. "Automating power: Social bot interference in global politics." *First Monday* 21 (4). http://firstmonday.org/ojs/index.php/fm/article/view/6161.

Wylie, C. 2020. *Mindf*ck: Inside Cambridge Analytica's Plot to Break the World*. London, UK: Profile Books.

ABOUT THE AUTHOR

Samantha Bradshaw is a post-doctoral fellow at the Digital Civil Society Lab and the Internet Observatory at Stanford University where she studies the relationship between technology and democracy, and the producers and drivers of disinformation and computational propaganda. Samantha's work has been published in leading academic journals, including *New Media & Society*, *Policy & Internet*, *Internet Policy Review* and the *Columbia Journal of International Affairs*. Her research and public writing have also been featured by international media including *The New York Times*, *The Washington Post* and CNN. Samantha completed her D.Phil. in information, communication and the social sciences at the Oxford Internet Institute, Oxford University.

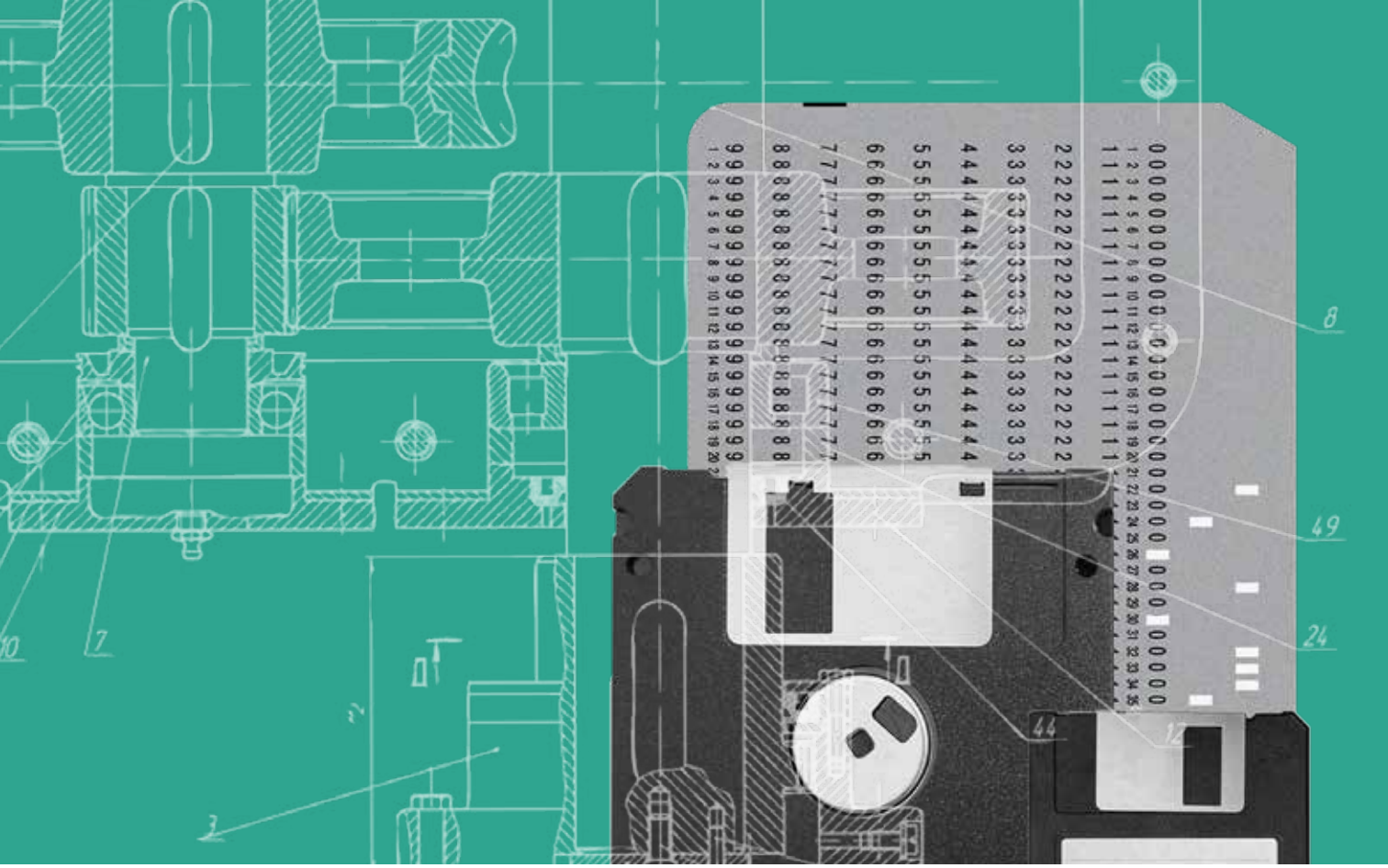
Subscribe

CIGI's weekly newsletter delivers the latest news, research, commentary and events to your inbox every Tuesday morning.

cigionline.org/subscribe

Artificial Intelligence and Keeping Humans "in the Loop"

Robert Mazzolin



Artificial intelligence (AI) technology has evolved through a number of developmental phases, from its beginnings in the 1950s to modern machine learning, expert systems and "neural networks" that mimic the structure of biological brains. AI now exceeds our performance in many activities once held to be too complex for any machine to master, such as the game Go and game shows. Nonetheless, human intellect still outperforms AI on many simple tasks, given AI's present inability to recognize more than schematic patterns in images and data. As AI evolves, the pivotal question will be to what degree AI systems should be granted autonomy, to take advantage of this power and precision, or remain subordinate to human scrutiny and supervision, to guard against unexpected failure. That is to say, as we anticipate technological advances in AI, to what degree must humans remain "in the loop"?

Computing is arriving at a critical juncture in its development. The traditional approaches relying on CMOS (complementary metal oxide semiconductor) technology, used in the manufacture of most of today's computer chips, and the pioneering architecture of John von Neumann are nearing their fundamental limits, and the speed of progress in computing power now seems to be falling short of the exponential improvement Moore's law would predict (Waldrop 2016). Further developments in the field of neuromorphic computing, in which semiconductors can imitate the structures of biological neurons and synapses, along with the advent of quantum computing, present a vision of human-level machine cognition serving as an intellectual partner to help solve some of the most significant technical, medical and scientific challenges confronting humankind.

Although AI researchers have had a checkered record in predicting the pace of technological progress, extrapolations of current trends suggest that AI with human-level cognition (artificial general intelligence) or above (artificial superintelligence) could be a relatively near-term prospect. Some experts predict an explosion in AI capabilities by 2045 (Baum, Goertzel and Goertzel 2011; Sandberg and Bostrom 2011), providing a massive supplement to the human brain, thereby dramatically increasing the general efficiency of human society. Such technology could grant a decisive strategic advantage in political, economic and military domains, and thus warrants the focused efforts of the world's leading nations.

Notwithstanding the current developmental challenges, there are technically no limits to the possible applications of AI.

As AI is now clearly being used in a comprehensive and world-changing way, a major challenge will be to make the processes and outputs of complex AI systems comprehensible to humans. This entails transparency of input data, algorithms and results that are clearly conveyed and easy to interpret. Enhanced transparency is a precondition for the acceptance of AI systems, particularly in mission-critical applications impacting life and death. Lack of user trust in AI decisions or understanding of how it functions will raise a host of legal, ethical and economic questions. The increasing delegation of human decisions to AI systems has varying consequences. Translation errors caused by automated systems such as Google Translate will likely have no serious impact on human life and survival. However, AI used in autonomous vehicles or weapon systems must make life-and-death decisions in real time. While it may be inconsequential to allow AI concerned with more mundane tasks to run without a human's finger hovering over the Off button, the use of AI technology to assist human cognition in more impactful decision making will likely require robust policies for retaining effective human control.

Notwithstanding the current developmental challenges, there are technically no limits to the possible applications of AI, which leads to ethical considerations. Emerging efforts focus on the development of AI technologies that can perceive, learn, plan, decide and act immediately in an environment of uncertainty. Some scholars predict an "intelligence explosion" beginning at the point in time when AI becomes more competent than humans at the very act of designing AI systems, setting AI development on an exponentially accelerating trajectory. This may lead to a "superintelligence," transcending the bounds of human thought, feeling and action. Such superintelligence could emancipate itself from human intelligence and arrive at different solutions than humans, given greater data, faster processing and, theoretically, more objective evaluation. The relative merit of such solutions may only be decided on the basis of values, raising the question of what canonical basis defines what is "right," and by whom, and whether by machine or not. These questions are particularly relevant in instances such as real-time conflict situations when human and machine values are incongruent and the competition for advantage in speed and accuracy may mean that humans will no longer be in charge, with those who refuse to delegate ultimate authority being outcompeted by those who do.

Merits of Humans "in the Loop" and "out of the Loop"

Given the prediction that future AI technology will be able to match or exceed human cognition across a wide range of tasks, the crucial question concerns the degree of autonomy that is most desirable. While AI often has the edge on humans in speed, efficiency and accuracy, its inability to think contextually and its tendency to fail catastrophically when presented with novel situations make many reticent to allow the technology to operate free of human oversight. As such, a common refrain is that a human must be kept in the loop to supervise AI in important roles. The AI would either have to attain a human supervisor's approval for its chosen course of action, or a human would

monitor an AI's actions, with the power to intervene should something go wrong.

It could be argued that removing human control could allow AI-enabled weaponry to temper the most distasteful elements of warfare and enable conflict to be conducted in an ethically superior manner. The intense stress, fatigue and emotional impulses endured by humans engaged in combat result in suboptimal decision making, frequently resulting in unnecessary collateral damage or unintended initiation of hostilities. The emotional and psychological causes that lie behind the accidental loss of human life during conflict cast doubt on the prospect of reforming human behaviour, but give reason for optimism that AI-enabled weapons could exceed human moral performance in similar circumstances (Arkin 2009). Consequently, one of the more attractive prospects of AI-enabled autonomous weapons is their imperviousness to such deficiencies, thereby enabling them to make more effective strategic decisions amid the "fog of war," or to kill in a more humane way (Lin, Bekey and Abney 2008). This argument potentially provides a strong case for the development and utilization of emotionally uncompromisable artificial combatants.

Conversely, the delegation of strategic decisions to AI could reduce the threshold for the onset of war, as machines would not be affected by the human mind's natural risk aversion. It could also cause armed conflicts to be prolonged endlessly, as machines do not tire or experience duress during extended periods of chaos and strife. Taking humans out of the loop and allowing autonomous weapon systems to operate fully independently complicates ethical and legal questions of liability and moral responsibility, such as the prosecuting of war crimes. Further, it is plausible that terrorist organizations could also adopt these technologies, possibly necessitating that lethal AI systems be deployed for peacetime policing activities as well.

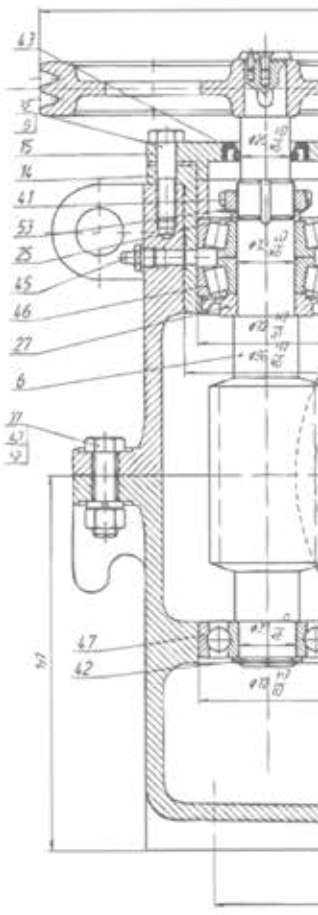
Finally, whether humans should be kept in the loop or not will depend upon how adept AI becomes at the crucial tasks of discriminating between different data sets to properly "self-learn," and noticing attempts at manipulation. At present, "data poisoning" and adversarial examples represent ways for malicious actors to exploit AI's inability to think contextually (Goodfellow et al. 2017). So long as this proves challenging for AI to overcome on its own, keeping a human overseer in the loop may be a necessary safeguard against such hostile actions.

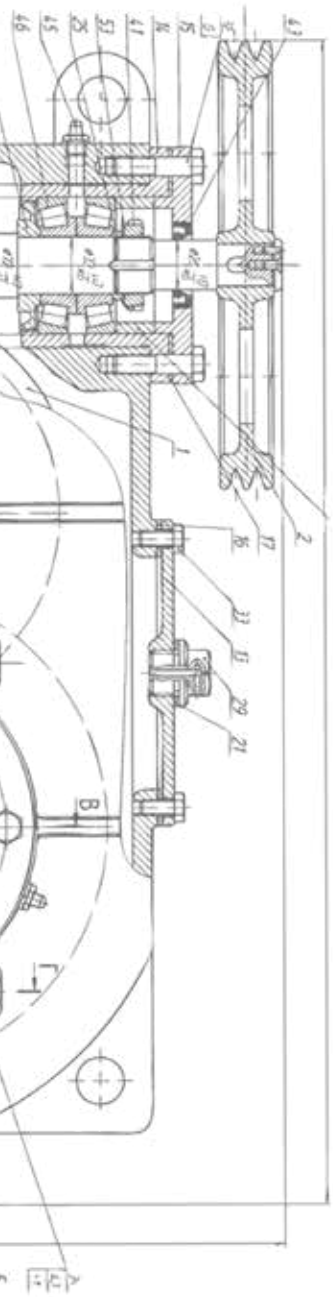
The Strategic Advantage of AI and Its Implications for Humans in the Loop

While different nations have a common interest in striking the right balance between autonomy and human supervision of AI, the fact that the multilateral global environment is being challenged by the re-emergence of great power competition will likely hinder international cooperation on this front.

The international race to develop advanced AI capabilities demonstrates the recognition by world leaders of the transformative potential of AI as a critical component of national security. The technology has the potential to change the international balance of power and to shape the course of unfolding geopolitical competition between the United States and China (and, to a lesser extent, Russia). To that end, each of these countries has implemented national initiatives that recognize the transformative effect that AI technology will have upon its security and strategic calculus. These states will be focused on maintaining information superiority, acquiring vast volumes of data to feed machine-learning algorithms. China's centralized planning, socialist market economy and the vast reservoir of data produced by its large population could give the country an advantage over competitors. Chinese policy has recently pushed for greater "civil-military fusion," seeking ways of adapting commercially developed technologies to the military sphere. President Xi Jinping has stated that AI, big data, cloud storage, cyberspace and quantum communications were among the "liveliest and most promising areas for civil-military fusion" (Chin 2018). The United States released its *National Artificial Intelligence Research and Development Strategic Plan* in 2016, and Russia reportedly harbours ambitions to make 30 percent of its force structure robotic by 2025 (National Science and Technology Council 2016; Eshel 2015).

The competing pursuit of AI technology by great and rising powers, as well as by non-state entities, promotes strategic competition, distrust and global instability. Societal dependence on the Internet of Things and





threats posed by AI-enabled cyberattacks will increase commensurately in both digital and physical domains, expanding the scope and scale of future cyberattacks. The many unexplainable elements of AI will compound these risks, further complicating security considerations in an uncertain and complex strategic landscape.

The growing intensity of this strategic competition may incentivize incautious policies toward human control of AI systems in military contexts. Speed is a crucial element of military effectiveness, and the ability of one actor to gather information, decide upon a course of action and execute these plans faster than its adversary has often proven key to victory. One of the most powerful advantages of AI systems is their ability to perform a given task much faster than a human. However, this advantage in speed may be undermined by efforts to keep humans in the loop. An autonomous weapon system that must prompt a human supervisor for approval before opening fire will be at a disadvantage against one that operates fully autonomously. At the pace at which AI systems are able to operate, the time lost on human decision making may prove the difference between victory and defeat. The aggressive Chinese and Russian pursuit of military-use AI and a relatively low moral, legal and ethical threshold in the use of lethal autonomous weapons may prompt the United States to shift from its current pledge to keep humans in the loop, which would intensify the emerging arms race in AI and adversely affect international security.

Ethical Issues and Meaningful Human Control

As AI systems are charged with making decisions with life-and-death consequences, be it in combat settings, medical facilities or simply on public roads, we are faced with the unpalatable prospect of dividing human lives into more and less valuable groups. Predictably, many are unsettled by the thought of a so-called "death algorithm," which takes this final decision independently. On what basis should an AI system determine which patient receives care when resources are stretched thin? What level of confidence must an AI weapon system have that a target is a combatant rather than a civilian before engaging?

Questions, including the following, arise:

- How much autonomy do societal consumers and decision makers wish to grant to AI technologies?
- What goals and purposeful manner will guide the establishment of ethical limits for AI's ability to make decisions that may impinge upon a target's fundamental rights and, ultimately, eliminate that target's life?
- More fundamentally, what moral framework does the decision maker utilize to decide?

Currently, there is no universally agreed-upon moral framework — divine command theory, utilitarianism and deontology represent various approaches. There is an element of subjectivity to these judgments, which is difficult, if not impossible, for current AI systems to satisfy. Therefore, international governance bodies should consider this issue seriously in the course of developing regulatory frameworks.

Governance Policy Development

Further scholarly work should be devoted to the unique challenges and risks posed by the need to exercise effective human oversight of increasingly complex AI systems. Analysis related to the societal impact of AI technology should include such topics as algorithmic transparency and the effects of AI on democracy. The prioritization and trade-offs between resource demands, accuracy, robustness and defence against attacks are other important considerations. Further, researchers need to consider potential mitigating measures, such as AI system patching to address software deficiencies as they arise in actual operation, and the application of "exit ramps" and "firebreaks," the programmatic decision points in the development of such systems to stop or amend the direction, scope and scale of activity to align with socially accepted standards.

China's astoundingly rapid developments in applying AI to an array of military applications demand close attention and scrutiny. As part of this critical examination, future research on AI-enabled weapon systems must account for the implicit values that are always embedded in the design of technologies and seek a governance framework that is both

precautionary and anticipatory. International governance bodies must understand the current limits of technology and become cognizant of how AI-enabled weapon systems are currently being developed regardless of societal concerns related to the nature and degree of human control. Consequently, national governments need to exercise caution when creating laws to govern the development and use of AI technologies, on account of the uncertainties that exist regarding how such laws will affect society. Moreover, governments must acknowledge the competitive pressure to remove human oversight of AI in military and security settings. It would be advisable to consider how prevailing knowledge surrounding effective arms-control agreements can be amended to suit the particular features of AI technology.

The concept of *meaningful human control* provides a helpful approach to discuss the employment, and ultimate weaponization, of increasingly autonomous AI technologies. This conceptual framework shifts the focus from speculation related to technological development and future capabilities toward the development and use of emerging technologies that conform with established societal norms related to responsibility, accountability, legality and humanitarian principles.

Finally, the AI science and engineering communities, represented via their professional societies, need to be engaged by governments and must articulate a position in the same manner as scientists in the areas of nuclear weapons, chemical agents and the use of disease agents in warfare. Active debate and position papers should be solicited as part of scientific societies' conferences and proceedings.

WORKS CITED

Arkin, Ronald C. 2009. "Ethical Robots in Warfare." *IEEE Technology and Society Magazine* 28 (1): 30–33. https://ieeexplore.ieee.org/document/4799405.

Baum, Seth D., Ben Goertzel and Ted G. Goertzel. 2011. "How Long Until Human-Level AI? Results from an Expert Assessment." *Technological Forecasting & Social Change* 78 (1): 185–95.

Chin, Josh. 2018. "China Looks to Close Technology Gap With U.S." *The Wall Street Journal*, April 21. www.wsj.com/articles/china-looks-to-close-technology-gap-with-u-s-1524316953.

Eshel, Tamir. 2015. "Russian Military to Test Combat Robots in 2016." *Defense Update*, December 31. http://defense-update.com/20151231_russian-combat-robots.html.

Goodfellow, Ian, Nicolas Papernot, Sandy Huang, Rocky Duan, Pieter Abbeel and Jack Clark. 2017. "Attacking Machine Learning with Adversarial Examples." *OpenAI* (blog), February 24. https://openai.com/blog/adversarial-example-research/.

Lin, Patrick, George A. Bekey and Keith Abney. 2008. *Autonomous Military Robotics: Risks, Ethics, and Design*. Investigative report, version 1.0.9, December 20. Washington, DC: US Department of Navy, Office of Naval Research. https://apps.dtic.mil/dtic/tr/fulltext/u2/a534697.pdf.

National Science and Technology Council. 2016. *The National Artificial Intelligence Research and Development Strategic Plan*. Washington, DC: Office of Science and Technology Policy, US Government. www.nitrd.gov/pubs/national_ai_rd_strategic_plan.pdf.

Sandberg, Anders and Nick Bostrom. 2011. "Machine Intelligence Survey." Technical Report #2011-1. Oxford, UK: Future of Humanity Institute, Oxford University. www.fhi.ox.ac.uk/reports/2011-1.pdf.

Waldrop, M. Mitchell. 2016. "The chips are down for Moore's law." *Nature*, February 9. www.nature.com/news/the-chips-are-down-for-moore-s-law-1.19338.

ABOUT THE AUTHOR

BGen (Retired) Robert Mazzolin (Ph.D., P.Eng., OMM, CD, SMIEEE) currently serves as the chief cybersecurity strategist for the RHEA Group, a space system engineering and cybersecurity organization delivering security solutions to large enterprises, governments and institutions in Europe and Canada. He retired from the Canadian Armed Forces (CAF) after serving as the vice director of Strategic Plans and Policy at United States Cyber Command at the National Security Agency in Fort Meade, Maryland. Notable appointments include Director General Information Management Operations, where he was responsible for all CAF and Department of National Defence strategic network, signals intelligence, electronic warfare and cyber operations; commander of the Canadian Forces Information Operations Group; director, Land Command Systems Program Management; commanding officer, Canadian Forces Station Leitrim and Canadian Forces Signals Intelligence Operations Centre. He served in a variety of other command and staff roles and was one of the Canadian Forces' leading experts in communications and information systems, signals intelligence, network operations and electronic warfare.

Centre for International
Governance Innovation

About CIGI

The Centre for International Governance Innovation (CIGI) is an independent, non-partisan think tank whose peer-reviewed research and trusted analysis influence policy makers to innovate. Our global network of multidisciplinary researchers and strategic partnerships provide policy solutions for the digital era with one goal: to improve people's lives everywhere. Headquartered in Waterloo, Canada, CIGI has received support from the Government of Canada, the Government of Ontario and founder Jim Balsillie.

À propos du CIGI

Le Centre pour l'innovation dans la gouvernance internationale (CIGI) est un groupe de réflexion indépendant et non partisan dont les recherches évaluées par des pairs et les analyses fiables incitent les décideurs à innover. Grâce à son réseau mondial de chercheurs pluridisciplinaires et de partenariats stratégiques, le CIGI offre des solutions politiques adaptées à l'ère numérique dans le seul but d'améliorer la vie des gens du monde entier. Le CIGI, dont le siège se trouve à Waterloo, au Canada, bénéficie du soutien du gouvernement du Canada, du gouvernement de l'Ontario et de son fondateur, Jim Balsillie.

cigionline.org

Autonomous systems are revolutionizing our lives, but they present clear international security concerns. Despite the risks, emerging technologies are increasingly applied as tools for cybersecurity and, in some cases, cyberwarfare. In this essay series, experts explore digital threats to democracy and security, and the geopolitical tensions they create.

cigionline.org/conflict-ai