Stockholm International Peace Research Institute

Report Part Title: Artificial intelligence: A primer
Report Part Author(s): VINCENT BOULANIN

Report Title: The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk
Report Subtitle: Volume I Euro-Atlantic Perspectives
Report Editor(s): VINCENT BOULANIN
Published by: Stockholm International Peace Research Institute (2019)
Stable URL: https://www.jstor.org/stable/resrep24525.8

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at https://about.jstor.org/terms



*Stockholm International Peace Research Institute* is collaborating with JSTOR to digitize, preserve and extend access to this content.

This content downloaded from 12.13.248.126 on Thu, 07 Sep 2023 20:39:28 +00:00
All use subject to https://about.jstor.org/terms

# Part I. Demystifying artificial intelligence and its military implications

This decade's renaissance of artificial intelligence (AI) has received growing attention from the popular media and policy community in recent years. However, despite the increasing number of publications and public events on the topic, widespread myths and misconceptions remain about what AI really is and what AI systems can do. This may seem surprising considering that this subfield of computer science has existed for more than half a century. The misconceptions are partly inherited from the way in which AI is depicted in popular culture, and in science fiction in particular. They are also reinforced by the way in which the media often tends to talk about technology: it uses eye-catching headlines that often exaggerate what the technology can do, for the best and for the worst. The problem with these misconceptions and misrepresentations is that they make it difficult—if not unproductive—to discuss the opportunities and risks of AI in general and in the security field in particular, as they obscure the true possibilities and limitations of AI technology. The following three essays thus aim to provide readers with a general but nuanced overview of what is currently happening in the field of AI.

The first essay (chapter 2) starts with a basic introduction to AI and discusses the components and implications of the current AI renaissance. Most notably, it explains why machine learning and autonomous systems are the focus of this volume. The subsequent essays provide the perspectives of a civilian engineer and a military engineer on the current state of AI technology, focusing on the case of autonomous systems. Dimitri Scheftelowitsch (in chapter 3) discusses in general terms what types of task state-of-the-art AI systems can and cannot undertake and why. Turning to the case of military systems, Martin Hagström (in chapter 4) puts the role of automation and autonomy in military systems in a historical perspective and assesses the role that machine learning could play in future military systems, including military autonomous systems.

VINCENT BOULANIN

This content downloaded from 12.13.248.126 on Thu, 07 Sep 2023 20:39:28 +00:00
All use subject to https://about.jstor.org/terms

This content downloaded from 12.13.248.126 on Thu, 07 Sep 2023 20:39:28 +00:00
All use subject to https://about.jstor.org/terms

# 2. Artificial intelligence: A primer

VINCENT BOULANIN

The concept of artificial intelligence (AI) was coined in the mid-1950s by John McCarthy, who defined it broadly as the 'science and engineering of making intelligent machines'.[1] However, the concept of AI means different things to different people, partly because its subject matter—intelligence—is hard to define.[2] Moreover, AI has, as one observer had put it, 'a bit of a Hollywood problem': popular culture has made it easy to talk about AI applications—from robots to smart digital assistant—but at the same time has 'skew[ed] expectations'.[3] There is a vast gap between the reality of what AI can do and the understanding, expectations and fears of the public, including often informed policymakers.

As a way to debunk potential misconception and misunderstanding around what AI is, this essay provides a basic introduction to this field of science and technology. It starts (in section I) by looking at how AI has been defined, how AI approaches have varied over time, and what is at stake with the current AI renaissance. It then presents two AI-related developments that arguably need particular attention when considering the potential of AI in the military sphere: machine learning (section II) and autonomous systems (section III).

## I. What is AI?

### The concept and its interpretations

For the majority of AI researchers, AI is about making machines capable of mimicking capabilities that are usually associated with human intelligence, such as observing the world through vision, processing natural language or learning. Some AI researchers differentiate between so-called narrow (or weak) AI and artificial general intelligence (AGI, or strong AI).

AGI is general-purpose AI: AI that would match—if not outperform—a human's ability to make sense of the world and to develop an understanding of its environment. It is the kind of AI that is typically depicted in popular culture in films such as *The Terminator*, *Blade Runner* or *2001: A Space Odyssey*. AGI has always fascinated AI researchers, but its design remains an unresolved technical challenge. There are, in fact, strong disagreements as to whether AGI will be ever

[1] Pearl, A., 'Homage to John McCarthy, the father of artificial intelligence (AI)', Artificial Solutions, 2 June 2017.

[2] Dale, R., 'An introduction to artificial intelligence', ed. A. M. Din, SIPRI, *Arms and Artificial Intelligence: Weapons and Arms Control Applications of Advanced Computing* (Oxford University Press: Oxford, 1987), p. 33.

[3] Madhavan, R., 'Understanding the societal impact of autonomous technologies', *IEEE Future Directions*, Nov. 2016.

This content downloaded from 12.13.248.126 on Thu, 07 Sep 2023 20:39:28 +00:00
All use subject to https://about.jstor.org/terms

be possible. Even the most optimistic AI researchers admit that AGI programs will remain in the realm of science fiction for the foreseeable future.[4]

Narrow AI has been around for decades and is the type of AI that is widely used today. Narrow AI systems are complex software programs that can execute discrete 'intelligent' tasks such as recognizing objects or people from images, translating language, or playing games. Narrow AI systems execute complex calculations, but they are brittle in nature—they are limited by the boundaries of their programming and they only work, at least reliably, for the intended tasks and operating environment. This is the type of AI that this volume refers to when it talks about AI.

For the purposes of this essay—and this volume more generally—AI is a catch-all term that refers to a wide set of computational techniques that allow computers and robots to solve complex, seemingly abstract problems that had previously yielded only to human cognition.[5] A key definitional element to bear in mind is that (narrow) AI is not a definite, unified technology. Instead, it is a portfolio technology that encompasses a wide variety of enabling applications which may be used to 'cognify' (i.e. give some form of cognitive capability to) multiple types of technology, including weapon technology.

AI is often depicted as a new or emerging technology, but it is not. As an academic discipline, it is more than half a century old. Narrow AI applications have been used for civilian and military purposes since the 1960s.[6]

A constant in the public debate has been the use of the concept of AI to refer to the newest computer technologies. Once these technologies have been widely deployed and adopted, they are no longer thought of or depicted as AI technologies. In other words, the frontier of AI is always moving: what is considered AI today may be considered normal software technology in the near future.

## The current AI hype: machine learning and autonomous systems

Since the 1950s, the field of AI has gone through several 'hype cycles': each period of major success (an 'AI summer') was inevitably followed by a period of disillusions (an 'AI winter') as the new and promising approach of AI eventually failed to match its early expectations.[7] These AI winters typically resulted in funding cutbacks.

Since the beginning of this decade, the field of AI has been experiencing a new summer due to a breakthrough in machine learning. This approach to AI

[4] Ready, C., 'Kurzweil claim that singularity will happen by 2045', Futurism, 5 Oct. 2017.

[5] International Panel on the Regulation of Autonomous Weapons (IPRAW), *Focus on Computational Methods in the Context of LAWS*, 'Focus on' Report no. 2 (German Institute for International and Security Affairs: Berlin, Nov. 2017).

[6] Dale (note 2); Russel, S. and Norvig, P., *Artificial Intelligence: A Modern Approach*, 3rd edn (Pearson Education: Harlow, 2014); Kit, P., 'What should we learn from past AI forecasts?', Open Philanthropy Project, May 2016; and Armstrong, S. and Sotala, K., 'How we're predicting AI—or failing to', eds J. Romportl et al., *Beyond AI: Artificial Dreams*, Proceedings of the International Conference 'Beyond AI 2012', Pilsen, Czechia, 5–6 Nov. 2012 (University of West Bohemia: Pilsen, 2012), pp. 52–75.

[7] On hype cycles see Gartner, 'Gartner hype cycle', [n.d.]; and Kit (note 6).

This content downloaded from 12.13.248.126 on Thu, 07 Sep 2023 20:39:28 +00:00
All use subject to https://about.jstor.org/terms

software development has been around since the beginning of AI research but has greatly benefited in the past decade from the progress of computer power and the increasing availability of digital data.[8]

Like previous AI summers, success stories about what current AI systems can achieve have channelled major interest and investment towards the most promising approach to AI engineering—currently machine learning—but also toward concrete applications that could be derived from it. One area of application that has received significant attention in recent years is autonomy. Autonomy (or 'machine autonomy') can be defined as the ability of a machine to execute a task or tasks without human input, using interactions of computer programming with the environment.[9] An autonomous system is, by extension, usually understood as a system—whether hardware or software—that, once activated, can perform some tasks or functions on its own. Progress in machine learning has opened major opportunities for increasing autonomous capabilities within systems and developing commercially viable autonomous systems, such as AI voice assistants, autonomous cars or autonomous weapons.

## II. Machine learning: A key enabler of the AI renaissance

### What is machine learning?

Machine learning is an approach to software development that first builds systems that can learn and then teaches them what to do using a variety of methods (i.e. supervised learning, reinforcement learning or unsupervised learning).[10] It removes the need for hand-coded programming, whereby humans hard-code software features into the systems.[11]

Machine learning was a marginal subfield of AI in the 1960s and 1970s as it was of limited practical use. In the 1980s and 1990s the digitalization of many industries and the emergence of large data sets—on which machine learning systems can be trained—reignited interest and inspired the development of new techniques. Among these were refined versions of a method known as 'artificial neural networks', which draws on knowledge of the human brain, statistics and applied mathematics.

The main advantage of machine learning compared to traditional hand-coded programming is that a human does not have to explicitly define the problem to be solved by the software and the way in which it solves the problem. Hand-coded programming usually requires a great deal of research on how the world works. In order to develop the model and rules that will govern the behaviour of an

[8] Knight, W., 'There is a big problem with AI', *MIT Technology Review*, 11 Apr. 2017.

[9] This definition is based on one previously proposed by Andrew Williams. Williams, A., 'Defining autonomy in systems: challenges and solutions', eds A. P. Williams and P. D. Scharre, *Autonomous Systems: Issues for Defence Policymakers* (NATO Headquarters Supreme Allied Commander Transformation: Norfolk, VA, 2015), pp. 27–62.

[10] For more details see Boulanin, V. and Verbruggen, M., *Mapping the Development of Autonomy in Weapon Systems* (SIPRI: Stockholm, Nov. 2017).

[11] Knight (note 8).

This content downloaded from 12.13.248.126 on Thu, 07 Sep 2023 20:39:28 +00:00
All use subject to https://about.jstor.org/terms

**Box 2.1.** Deep learning

Deep learning is an approach to machine learning whereby the system 'learns' how to undertake a task in supervised, semi-supervised or unsupervised ways. It transforms raw data input into abstract representations (features) that can be effectively exploited in machine learning tasks, such as recognizing an object in an image.

The strength of deep learning is in its ability to introduce representations that are expressed in terms of other, simpler representations. In other words, it allows the computer to build complex concepts from simpler concepts. A deep-learning system can, for instance, represent the concept of an image of a person by combining simple concepts, such as corners and contours.[a]

The success of deep learning was supported by two trends. First was the widespread commercialization of graphics processing units (GPUs), a type of computer chip that is well suited for machine learning operations. Second, and perhaps more importantly, was the widespread democratization of the Internet and social media, which led to an explosion in the volumes of the digital data on which machine learning algorithms can be trained.

In recent years, deep learning has become the most fashionable approach to artificial intelligence (AI) engineering, but this does not mean that it has totally supplanted other approaches. Many cutting-edge AI applications do not use deep learning—or even machine learning. Many continue to rely on relatively old-fashioned hard-coded expert knowledge, while others may use established machine learning methods such as Bayesian statistics or evolutionary algorithms.[b] Traditional AI programming methods will, in other words, continue to be relevant. In fact, some argue that truly intelligent machines can only be develop by combining machine learning with traditional programming that can introduce abstract knowledge and a layer of common sense reasoning.[c]

[a] Goodfellow, I., Bengio, Y. and Courville, A., *Deep Learning* (MIT Press: Cambridge, MA, 2016), p. 8.

[b] Hao, K., 'We analyzed 16,225 papers to figure out where AI is headed next', *MIT Technology Review*, 25 Jan. 2019.

[c] Thompson, C., 'How to teach artificial intelligence some common sense', *Wired*, 13 Nov. 2018.

autonomous system, the developing engineers often cooperate with scientists from other scientific fields, notably the natural sciences (e.g. neurosciences and physics) and the social sciences (e.g. psychology, linguistics and sociology). Hand-coded programming gets difficult for tasks and operating environments too complex for a human to model completely.[12] One such problem is image recognition. It is hard for human to express in mathematical terms what distinguishes pictures of cats from pictures of dogs. Machine learning allows engineers to bypass this difficulty, as it allows an image-recognition system to generate its own way of perceiving the differences.[13]

When used in non-technical context, the term 'learning' can sometimes be a source of confusion, as it invites an anthropomorphic interpretation. However, the way in which machine learning works has nothing to do with the way humans learn: machines learn by finding statistical relationships in past data.[14] Engineers use the term 'learning' for practical reasons: it is a concise and memorable way of describing a complex computing process.

[12] Kester, L., 'Mapping autonomy', Presentation at the CCW Informal Meeting of Experts on Lethal Autonomous Weapon Systems, Geneva, 11–15 Apr. 2016.

[13] Knight (note 8).

[14] Boulanin and Verbruggen (note 10).

This content downloaded from 12.13.248.126 on Thu, 07 Sep 2023 20:39:28 +00:00
All use subject to https://about.jstor.org/terms

## The rise of deep learning as the dominant approach to AI engineering

In 2006 Geoffrey Hinton, a cognitive psychologist and computer scientist at the University of Toronto, co-authored an academic paper on what was deemed at the time to be a hopelessly musty academic problem in the AI community: how to make neural networks work more quickly and more effectively.[15] A neural network is a possible architecture for machine learning that was briefly popular in the 1980s but then became marginal and unfunded: reportedly, by the early 2000s, no more than five researchers specialized in neural networks.[16] Hinton was one of the last scholars in the AI community who believed that machine learning, and neural networks in particular, was a promising approach for the development of truly proficient AI systems. The paper demonstrated that, when combined in different layers, neural networks could be very powerful. When they published it, Hinton and his co-authors did not realize that they were onto something that was poised to not only transform the field of AI from the inside but also to reignite massive interest in machine learning and the field of AI more generally.[17]

The technique they came up with—later rebranded 'deep learning'—was to outperform traditional AI programming techniques by a wide margin (see box 2.1). AI systems that would rely on deep learning systematically beat existing AI systems at implementing tasks such as recognizing images or speech, translating language, or playing games such as chess.[18]

Industry and governments alike soon identified major opportunities and started to invest in deep learning and machine learning more broadly. Large technology companies such as Google and Facebook were the fastest to react. At the beginning of this decade, they bought the most innovative start-up companies that worked with deep learning.[19] They also recruited the few scholars that had remained active in this field and created dedicated research teams.[20] Many applications that these companies commercialize today—from smartphones to social media platforms—are powered by deep learning.

Governments also gradually developed a major interest in deep learning and the promise that it held in terms of enabling the development and use of practical and powerful applications in both the civilian and military spheres. Governmental research funding institutions, such as the United States' Defense Advanced Research Projects Agency (DARPA), started pouring money into machine learning-related research and development projects, while governments started

[15] Hinton, G. E., Osindero, S. and Teh, Y., 'A fast learning algorithm for deep belief nets', *Neural Computation*, vol. 18, no. 7 (July 2006), pp. 1527–54. See also Kurenkov, A., 'A "brief" history of neural nets and deep learning, part 4', *Medium*, 17 Feb. 2017.

[16] Allen, K., 'How a Toronto professor's research revolutionalized artificial intelligence', *Toronto Star*, 17 Apr. 2015.

[17] Somers, J., 'Is AI riding a one-trick pony', *MIT Technology Review*, 29 Sep. 2017.

[18] Gershgorn, D., 'The data that transformed AI research—and possibility the world', *Quartz*, 26 July 2017.

[19] E.g. Shu, C., 'Google acquires AI startup DeepMind for more than $500m', TechCrunch, 26 Jan. 2014.

[20] Malingo, T., 'The misguided rush of the academic AI brain drain', New Stack, 23 Aug. 2018.

This content downloaded from 12.13.248.126 on Thu, 07 Sep 2023 20:39:28 +00:00
All use subject to https://about.jstor.org/terms

**Box 2.2.** Generative adversarial networks

The generative adversarial network (GAN) is a new approach (invented in 2014) that involves two artificial neural network systems that can spar with each other to create ultra-realistic original image, audio or video content—something that machines have never been able to do properly before.

The two networks are trained on the same data set. One, known as the generator, is tasked with creating variations on images that it has already seen. The second, known as the discriminator, is asked to identify whether the example it sees is like the images on which it has been trained or a fake produced by the generator. Over time, the generator becomes so good that the discriminator cannot spot fakes.

The potential consequences of this breakthrough are both positive and negative. On the one hand, it could help a machine learning system to generate new data to train itself; on the other hand, it could create digital fakery for criminal or information warfare purposes.

*Source*: Condliffe, J., 'Duelling neural networks: by playing cat-and-mouse games with data, a pair of AI systems can acquire an imagination', *MIT Technology Review*, vol. 121, no. 2 (Mar./Apr. 2018).

internal discussions on what their strategy towards deep learning should be.[21] Between 2017 and early 2019 at least 17 countries released a national strategy or made a strategic policy announcement on AI.[22]

## Opportunities and challenges

Whether from the industry or the governmental side, there is a common understanding that the AI renaissance that Hinton and his peers initiated has brought its share of opportunities and challenges.[23]

On the opportunity side, machine learning no longer needs to demonstrate its potential. It has already dramatically improved the ability of computers and robots to perceive the world, which has accelerated the development of autonomous systems such as self-driving cars and voice assistants.[24] Machine learning has also shown itself to be a powerful tool for data management. It can be used not only to classify data but also to find correlations in data that can then be used to make statistical predictions about future behaviour. Internet service providers, such as Google, Facebook or Baidu, routinely use machine learning to label and organize content such as text, images and videos, and to predict customer preferences.[25] National militaries are now trying to develop a similar capability to process

[21] US Defense Advanced Research Project Agency (DARPA), 'DARPA announce $2 billion campaign to develop next wave of AI technologies', 7 Sep. 2018.

[22] These 17 countries are Canada (Mar. 2017), Japan (Mar. 2017), Singapore (May 2017), China (July 2017), the United Arab Emirates (Oct. 2017), Finland (Dec. 2017), Denmark (Jan. 2018), Italy (Mar. 2018), France (Mar. 2018), the UK (Apr. 2018), South Korea (May 2018); Sweden (May 2018), India (June 2018), Mexico (June 2018), Germany (Nov. 2018), the European Union (Dec. 2018) and the USA (Feb. 2019). Dutton, T., 'An overview of national AI strategies', Medium, 18 June 2018.

[23] As identified by Scharre, P. and Horowitz, M. C., *Artificial Intelligence: What Every Policymaker Needs to Know* (Center for New American Security: Washington DC, June 2018), p. 10.

[24] Gershgorn, D., 'See the difference one year makes in artificial intelligence research', *Popular Science*, 31 May 2016.

[25] Marr, B., 'The amazing way Google uses deep learning AI', *Forbes*, 8 Aug. 2017.

This content downloaded from 12.13.248.126 on Thu, 07 Sep 2023 20:39:28 +00:00
All use subject to https://about.jstor.org/terms

intelligence data.[26] The ability of machine learning to identify patterns can also be used to detect anomalies in data. Cybersecurity companies increasingly rely on machine learning to detect new malware, while the surveillance industry is exploring the possibility of using machine learning to detect abnormal behaviour in CCTV footage.[27] Machine learning has also turned out to be useful for optimizing the performance of complex tasks, such as controlling large numbers of robots.[28] It can even generate new content such as original, ultra-realistic images, sounds or written stories. This has been achieved with a new machine learning-based approach known as a generative adversarial network (GAN, see box 2.2), which could reportedly become the next big thing in the AI community.[29] Machine learning holds great promise, but it also has significant shortcomings. The first—and perhaps most salient—relates to its dependence on data. Systems that are powered by machine learning are only as good as the data on which they are trained.[30] To be taught, a machine learning system needs to be provided with large volumes of real word examples (training data) and rules about the data relationships. In order to recognize a type of object in an image (e.g. a car, a bus or a dog), a computer vision system may need to be trained with millions of pictures of that type of object. The quality of the data on which the systems are trained is equally important. If the training data set is not representative, then the system might fail, might perform poorly, or might misinform human decisions and actions by reinforcing existing human biases or creating new ones.[31] Research has shown, for instance, that facial recognition systems trained with data sets that primarily include images of light-skinned men are more likely to misidentify faces of women or people with darker skin.[32] Companies and governments willing to develop and use machine learning therefore have a double challenge: they have to both find a sufficiently large set of training data and a way to ensure that this data is reliable. A number of recent news reports indicates that companies and governments alike are struggling with that challenge.[33]

Another fundamental, and related, problem is the fact the machine learning algorithm are opaque. Traditional handcrafted AI systems reason according to rules and logic, making the inner workings transparent to anyone who cares to

[26] One such system—the USA's Project Maven—is discussed in chapters 5, 6, 10 and 11 in this volume.

[27] Polyakov, A., 'Machine learning for cybersecurity 101', Towards Data Science, 4 Oct. 2018.

[28] Hüttenrauch, M., 'Guided deep reinforcement learning for robot swarms', Master's thesis, Technische Universität Darmstadt, Aug. 2016.

[29] Condliffe, J., 'Duelling neural networks: by playing cat-and-mouse games with data, a pair of AI systems can acquire an imagination', *MIT Technology Review*, vol. 121, no. 2 (Mar./Apr. 2018).

[30] Gershgorn (note 18); and Hao, K., 'We analyzed 16,225 papers to figure out where AI is headed next', *MIT Technology Review*, 25 Jan. 2019.

[31] Knight, W. and Hao, K., 'Never mind killer robots—here are six real AI dangers to watch out for in 2019', *MIT Technology Review*, 7 Jan. 2019; and Hao, K., 'This is how AI bias really happens—and why it's so hard to fix', *MIT Technology Review*, 4 Feb. 2019.

[32] Lohr, S., 'Facial recognition is accurate, if you're a white guy', *New York Times*, 9 Feb. 2018.

[33] See e.g. Buranyi, S., 'Rise of the racist robots—how AI is learning all our worst impulses', *The Guardian*, 8 Aug. 2017; Hao, K., 'Police across the US are training crime-predicting AIs on falsified data', *MIT Technology Review*, 13 Feb. 2019; and Hao, K., 'AI is sending people to jail—and getting it wrong', *MIT Technology Review*, 21 Jan. 2019.

This content downloaded from 12.13.248.126 on Thu, 07 Sep 2023 20:39:28 +00:00
All use subject to https://about.jstor.org/terms

examine the code. In contrast, a machine learning system, particularly one that relies on deep neural networks, operates like a black box.[34] The input and the output of such a system are observable, but the computational process leading from one to the other is difficult for humans to understand. It is particularly difficult for humans to understand what such a system has learned and hence how it might react to input data that is different from that used during the training phase.[35] The lack of transparency and explainability of these systems in turn creates a fundamental problem of predictability. A machine learning system might fail in ways that were unthinkable to humans because the engineers do not have a full understanding of its inner working. In the context of weapon systems, this unpredictability could have dramatic consequences. The lack of transparency is also problematic from a regulatory standpoint as it makes complex the task of identifying the source of a problem and attributing responsibility when something goes wrong.[36]

Moreover, AI systems trained with machine learning may outperform human for many tasks, but they still lack what humans understand as basic common sense. Computer vision systems, for instance, do not perceive a pattern at an abstract level, like a human would. They just see a correlation between a group of pixels. A facial recognition system could not tell the difference between an actual person and a picture of a picture within an image: in both cases it would be a positive identification. One study also recently demonstrated that variations in an image that are imperceptible to the human eye could cause an image-recognition system to completely mislabel the object or people in the image (e.g. mistaking a lion for a car or a building).[37] Another study has demonstrated that it is easy to produce images that are completely unrecognizable to humans but that computer vision software believes to be a recognizable object with over 99 per cent confidence.[38] In other words, machine learning systems may not be reliable: they can easily be fooled (which is particularly problematic if used in an adversarial context such as an armed conflict) or they may fail in unpredictable ways according to human standards.

In sum, while recent advances in machine learning have created important opportunities for the development of highly efficient AI systems, including autonomous systems, machine learning is still, in several regards, an immature technology. There remain many technical practical challenges associated with the use of machine learning methods. The fundamental question that developers, users and regulators are currently struggling with is how to ensure the responsible

[34] Knight, W., 'The dark secret at the heard of AI', *MIT Technology Review*, 11 Apr. 2017.

[35] Righetti, L., 'Emerging technology and future autonomous systems', International Committee of the Red Cross (ICRC), *Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons*, Expert meeting, Versoix, Switzerland, 15–16 Mar. 2016 (ICRC: Geneva, Aug. 2016), pp. 36–39.

[36] Tobey, D., 'Explainability: where AI and liability meet', DLA Piper, 25 Feb. 2019.

[37] Szegedy, C. et al., 'Intriguing properties of neural networks', arXiv, 1312.6199, version 4, 19 Feb. 2014.

[38] Nguyen, A., Yosinski, J. and Clune J., 'Deep neural networks are easily fooled: high confidence predictions for unrecognizable images', *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, Proceedings, 7–12 June 2015 (Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, 2015), pp. 427–36.

This content downloaded from 12.13.248.126 on Thu, 07 Sep 2023 20:39:28 +00:00
All use subject to https://about.jstor.org/terms

adoption and use of this technology. This question is particularly pressing for systems and applications that are safety critical (e.g. cars, aeroplanes and weapons) or that could have a societal impact (e.g. health and educational services).

## III. Autonomy: A key by-product of the AI renaissance

### What is autonomy?

As a technology area, autonomy is related to but distinct from AI. While machine learning can be depicted as the key ingredient of the current renaissance of AI (and the associated hype), autonomy can be portrayed as one of its key by-products. Autonomous systems ranging from AI assistants (e.g. Amazon's Alexa or Apple's Siri), via self-driving cars and auto-piloted unmanned aerial vehicles (UAVs) to autonomous weapons are among the most debated technological developments deriving from the current AI renaissance and with the highest level of media attention.[39]

As discussed above, 'autonomy' can be understood as the ability of a machine to execute a task, or tasks, without human input, using interactions of sensors, computer programming and actuators with the environment.[40] However, the concept of autonomy and autonomous systems more generally mean different things to different people, primarily because autonomy is a relative notion that can be interpreted in several ways. The level of autonomy of a system can be analysed from three different and independent perspectives (see figure 2.1): (*a*) based on the extent to which humans are involved in the execution of the task carried out by the system; (*b*) based on the extent to which the system can exercise control over its own behaviour and deal with uncertainties in its operating environment; and (*c*) based on the number and types of functions that are automated.[41]

This definitional uncertainty appears clearly in this volume: the contributors sometimes use different terminology and metrics to talk about autonomy. For example, some experts make a distinction between automatic, automated and autonomous systems, while others use these terms interchangeably (see box 2.3). For the analytical purposes of this volume, it is not necessary to resolve that conceptual debate: the level of autonomy of systems can be analysed through different lenses and each lens has its own analytical value.

[39] Hao, K., 'One day your voice will control all your gadgets, and they will control you', *MIT Technology Review*, 11 Jan. 2019; 'Autonomous weapon and the new laws of war', *The Economist*, 17 Jan. 2019; and Salesky, B., 'A decade after DARPA: our view on the state of the art in self-driving cars', *Medium*, 16 Oct. 2017.

[40] Williams (note 9), pp. 55–56.

[41] As identified by Scharre, P., 'The opportunity and challenge of autonomous systems', eds Williams and Scharre (note 9), pp. 3–26. See also Thrun, S., 'Toward a framework for human-robot interaction', *Human–Computer Interaction*, vol. 19, nos 1–2 (June 2004), pp. 9–24; and Boulanin and Verbruggen (note 10).

This content downloaded from 12.13.248.126 on Thu, 07 Sep 2023 20:39:28 +00:00
All use subject to https://about.jstor.org/terms

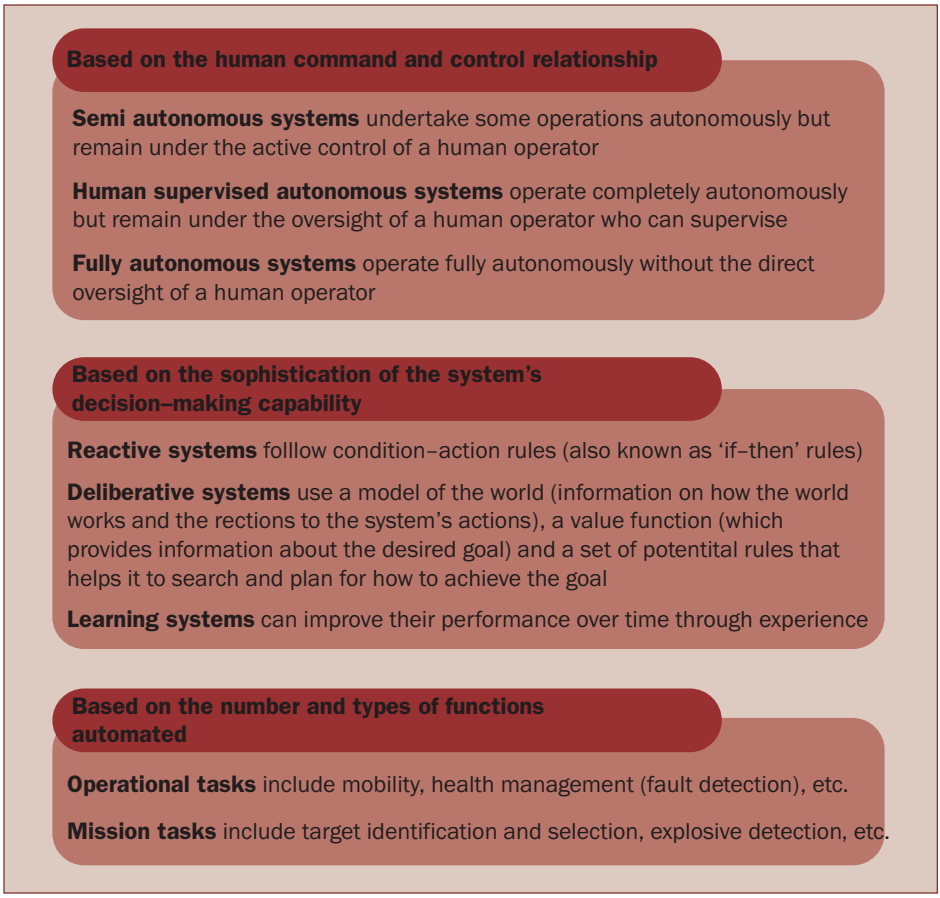



**Figure 2.1.** Approaches to the definition and categorization of autonomous systems

*Source*: Boulanin, V. and Verbruggen, M., *Mapping the Development of Autonomy in Weapon Systems* (SIPRI: Stockholm, Nov. 2017).

## Opportunities and challenges[42]

Advances in autonomy are generating great expectations in both the civilian and military spheres as they enhance the usefulness and reliability of robotics systems, which in turn could generate significant economic and operational benefits. Companies, governmental institutions and the military alike could achieve greater manpower efficiency by increasing their reliance on robotic systems.[43] Advances in autonomy could also allow them to overcome a number of operational challenges associated with manned operations or the use of teleoperated systems.[44]

[42] This section is based on Boulanin and Verbruggen (note 10), pp. 61–82.

[43] US Department of Defense (DOD), Defense Science Board, *The Role of Autonomy in DoD Systems*, Task Force Report (DOD: Washington, DC, July 2012); and Scharre, P., *Robotics on the Battlefield*, part II, *The Coming Swarm* (Center for New American Security: Washington, DC, Oct. 2014).

[44] eds Williams and Scharre (note 9).

This content downloaded from 12.13.248.126 on Thu, 07 Sep 2023 20:39:28 +00:00
All use subject to https://about.jstor.org/terms

**Box 2.3.** Automatic, automated, autonomous

*Automatic*

The label 'automatic' is usually reserved for systems that mechanically respond to sensory input and step through predefined procedures, and whose functioning cannot accommodate uncertainties in the operating environment. An example of this is a robotic arm used in the manufacturing industry.

*Automated versus autonomous*

Machines that can cope with variations in their environment and exercise control over their actions can be described as either automated or autonomous. What distinguishes an automated system from an autonomous system is a contentious issue.

Some experts see the difference in terms of the degree of self-governance. They view autonomous systems merely as more complex and intelligent forms of automated system.

Others see value in making a clear distinction between the two concepts. A report from the US Defense Science Board presents an automated system as a system that is governed by 'prescriptive rules that permit no deviations'.[a] This means that the system logically follows a pre-defined set of rules in order to provide an outcome; its output is predictable if the set of rules under which it operates is known. In contrast, an autonomous system is able to 'independently compose and select among different courses of action to accomplish goals based on its knowledge and understanding of the world, itself, and the situation'.[b]

While the distinction between the terms automatic, automated and autonomous can be conceptually useful, in practice it is difficult to determine which of the three categories a system belongs to. Moreover, the definitions of and boundaries between these three categories remain contested within and between the expert communities.

[a] US Department of Defense (DOD), Defense Science Board, *Report of the Defense Science Board Summer Study on Autonomy* (DOD: Washington, DC, June 2016), p. 4.

[b] US Department of Defense (note a), p. 4.

*Source*: Boulanin, V. and Verbruggen, M., *Mapping the Development of Autonomy in Weapon Systems* (SIPRI: Stockholm, Nov. 2017).

Autonomy means that a robotic system can execute some tasks much faster than any human or human-operated robot ever could, which for the military is particularly attractive for time-critical missions or tasks such as air defence, air-to-air combat or cyber-defence. Autonomy can make robotic systems far more agile from a command-and-control perspective and reduce the need to maintain a constant communications link between the robot and the military command. It can also allow the military to reduce the number of human operators and analysts needed to oversee the system and process information. Autonomy is also useful for so-called dull, dirty and dangerous (3D) missions as it removes limitations (e.g. fatigue, boredom, hunger or fear) that may make human performance deteriorate over time. Autonomy also gives systems greater reach. It grants access to operational theatres that were previously inaccessible to remote-controlled systems or too risky for manned operations. These include areas protected by anti-access/area-denial (A2/AD) systems and areas with harsh operating environments for humans (and where communication is limited), such as deep water, the Arctic and, potentially, outer space. Finally, autonomy also provides new opportunities for collaborative operations as it permits weapon systems to

This content downloaded from 12.13.248.126 on Thu, 07 Sep 2023 20:39:28 +00:00
All use subject to https://about.jstor.org/terms

operate in large groups, or 'swarms', in a much more coordinated, structured and strategic way than if they were individually controlled by a human operator.

The advances in autonomy are, however, raising a wide spectrum of ethical, legal and security concerns, which apply to both civilian and military applications.[45] From an ethical standpoint, the development of autonomy in safety-critical systems such as cars or weapons raises the vexed question of whether, and to what extent, autonomous systems should be trusted to operate outside of direct human control and supervision. The ongoing intergovernmental discussion on lethal autonomous weapon systems (LAWS) in the framework of the 1980 Convention on Certain Conventional Weapons (CCW Convention) and the debate around car accidents involving semi-autonomous cars have shown that there is no simple answer to that question.[46] The question of the balance between autonomy and human control also has profound and complex legal implications, particularly with regard to the attribution of individual criminal responsibility: who should be prosecuted when a self-driving car or autonomous weapon causes harm? Some legal experts argue that the proliferation of autonomous systems is taking place in a legal vacuum or grey zone, which could make the determination of legal accountability difficult in case of a deadly incident.[47] Advances in autonomy also create new security risks. In addition to the increasing vulnerability to cyberattacks, the limitations of existing autonomous systems in terms of perceptual and decision-making intelligence could easily be exploited by a malevolent actor who could defeat a system by simply spoofing the sensors or control systems.[48]

More broadly, the increasing adoption of and reliance on autonomous systems is bound to ignite profound societal changes.[49] Among other effects, it will change the way that companies, governmental agencies and the military operate. Taking the case of an air force as an example, replacing manned combat aircraft with autonomous unmanned aerial systems will necessitate a change in the way that personnel are selected, trained and meant to operate—with control moving from a pilot to a remote operator and then to a systems supervisor—which in turn could cause major changes in professional culture.[50]

[45] Cath, C. et al., 'Governing artificial intelligence: ethical, legal and technical opportunities and challenges', *Philosophical Transactions of the Royal Society A*, vol. 376, no. 2133 (Nov. 2018).

[46] Boulanin, V., 'Mapping the debate on LAWS at the CCW: taking stock and moving forward', EU Non-proliferation Paper no. 49, EU Non-proliferation Consortium, Mar. 2016; Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons which may be Deemed to be Excessively Injurious or to have Indiscriminate Effects (CCW Convention), opened for signature 10 Apr. 1981, entered into force 2 Dec. 1983; and Bhuyian, J., 'Uber's semi-autonomous cars detected the pedestrian six seconds before the fatal crash, a federal agency says', Recode, 25 May 2018.

[47] Docherty, B., *Mind the Gap: The Lack of Accountability for Killer Robots* (Human Rights Watch: New York, 2015).

[48] Versprille, A., 'Army still determining the best use for driverless vehicles', *National Defense*, June 2015; and Endsley, M. R., *Autonomous Horizons: System Autonomy in the Air Force—A Path to the Future*, vol. 1, *Human–Autonomy Teaming* (US Air Force, Office of the Chief Scientist: Washington, DC, 2015), p. 5.

[49] Wright, N., 'Three distinct AI challenges for the UN', AI & Global Governance, United Nations University, Centre for Policy Research, 7 Dec. 2018.

[50] For detailed discussion see Boulanin and Verbruggen (note 10), pp. 69–73; and chapter 12 in this volume.

This content downloaded from 12.13.248.126 on Thu, 07 Sep 2023 20:39:28 +00:00
All use subject to https://about.jstor.org/terms

## IV. Conclusions

In a June 2018 essay in *The Atlantic*, Henry Kissinger, a former US Secretary of State, argues that 'The Enlightenment started with essentially philosophical insights spread by a new technology. Our period is moving in the opposite direction. It has generated a potentially dominating technology in search of a guiding philosophy.'[51] The dominating technology that he refers to is artificial intelligence. Recent advances in machine learning have unlocked numerous possibilities, including that of creating increasingly autonomous systems. However, developers and users alike have only begun to work out how to make the best use of it, and also how they should not use it.[52]

Big AI companies have come together and launched several initiatives that are meant to lead the global reflection on these questions, notably the Partnership on AI and OpenAI.[53] The Institute of Electrical and Electronics Engineers (IEEE), the world's largest association of engineers, has also started a conversation, which is not limited to its members, on the responsible development of AI systems.[54] Some states, through a national strategy on AI, are trying to work out a national approach to these issues, and also to identify concrete solutions. For example, the French national strategy on AI calls for funding of research and development work that would make machine learning systems more explainable in order to make them more socially acceptable.[55] It also supports the inclusion of ethics in training for AI engineers and researchers and the implementation of social impact assessments on new machine learning systems.

So far, most of these efforts have been connected to the civilian development of AI technology. The debate on what should be the norms and principles for responsible development of AI systems in the military sphere has gained momentum but has been constrained by the framework of the ongoing CCW debate on LAWS, which, of course, focuses on conventional weapons. The remainder of this volume opens up that conversation to the field of nuclear weapons and doctrines. The essays that follow are intended to lay the foundations for a constructive conversation on the possibilities and challenges that the current AI renaissance will bring to the realm of nuclear weapons and doctrines and to help identify possible norms and principles for the responsible development of AI systems in that context.

[51] Kissinger, H., 'How the enlightenment ends', *The Atlantic*, June 2018.

[52] Vogt, H., 'Artificial intelligence rules more of your life. Who rules AI?', *Wall Street Journal*, 13 Mar. 2018; and Hao, K., 'Why AI is a threat to democracy—and what we can do to stop it', *MIT Technology Review*, 26 Feb. 2019.

[53] Partnership on AI, 'About us', [n.d.].

[54] Madhavan (note 3); IEEE Standards Association, 'The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems', [n.d.]; and IEEE Global Initiative, *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, version 2 (Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, [Dec. 2017]).

[55] Villani, C., *For a Meaningful Artificial Intelligence: Toward a French and European Strategy*, (Conseil national du numérique: Paris, Mar. 2018).

This content downloaded from 12.13.248.126 on Thu, 07 Sep 2023 20:39:28 +00:00
All use subject to https://about.jstor.org/terms