UNIDIR

The Weaponization of Increasingly Autonomous Technologies: Artificial Intelligence

a primer for CCW delegates

No. 8 UNIDIR RESOURCES

Acknowledgements

Support from UNIDIR's core funders provides the foundation for all of the Institute's activities.

UNIDIR would like to thank those who provided input to this paper, in particular Paul Scharre, as well as those who kindly reviewed it at various stages of its production.

About the Project "The Weaponization of Increasingly Autonomous Technologies"

Given that governments have a responsibility to create or affirm sound policies about which uses of autonomy in weapon systems are legitimate—and that advances in relevant technologies are also creating pressure to do so—UNIDIR's work in this area is focused on what is important for States to consider when establishing policy relating to the weaponization of increasingly autonomous technologies. See http://bit.ly/UNIDIR_Autonomy for Observation Papers, audio files from public events, and other materials.

This is the eighth in a series of UNIDIR papers on the weaponization of increasingly autonomous technologies. UNIDIR has purposefully chosen to use the word "technologies" in order to encompass the broadest relevant categorization. In this paper, this categorization includes machines (inclusive of robots and weapons) and systems of machines (such as weapon systems), as well as the knowledge practices for designing, organizing and operating them.

About UNIDIR

The United Nations Institute for Disarmament Research—an autonomous institute within the United Nations—conducts research on disarmament and security. UNIDIR is based in Geneva, Switzerland, the centre for bilateral and multilateral disarmament and non-proliferation negotiations, and home of the Conference on Disarmament. The Institute explores current issues pertaining to the variety of existing and future armaments, as well as global diplomacy and local tensions and conflicts. Working with researchers, diplomats, government officials, NGOs and other institutions since 1980, UNIDIR acts as a bridge between the research community and governments. UNIDIR's activities are funded by contributions from governments and foundations.

Note

The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations concerning the legal status of any country, territory, city or area, or of its authorities, or concerning the delimitation of its frontiers or boundaries.

The views expressed in this publication are the sole responsibility of UNIDIR. They do not necessarily reflect the views or opinions of the United Nations or UNIDIR's sponsors.

@UNIDIR
www.unidir.org

Contents

Acronyms and abbreviations

AGI	Artificial General Intelligence
AI	Artificial Intelligence
ASIC	Application-Specific Integrated Circuit
CCW	Convention of Certain Conventional Weapons
GGE	Group of Governmental Experts
LAWS	Lethal Autonomous Weapon Systems
TPU	Tensor Processing Units

Introduction

Since governments began international discussions on lethal autonomous weapon systems (LAWS) in 2014,[1] the field of artificial intelligence (AI) has seen tremendous advances. In just the past few years, AI systems have surpassed human abilities in benchmark games such as poker and Go and have begun to be applied to a range of real-world problems in medicine, finance, transportation, and other industries.

The rapidly advancing field of AI and machine learning has significant implications for the role of autonomy in weapon systems. More intelligent machines are capable of taking on more challenging tasks in more complex environments. Some tasks that machines performed poorly at (relative to humans) in 2014, such as image recognition, machines now can reliably perform better than humans.[2] This complicates discussions surrounding the role of autonomy in weapons. Because of the rapid progress in AI, participants in international discussions may have different perspectives on what is possible today, to say nothing of their expectations about what may be possible in the next few years.

For example, only a few weeks before discussions at the Group of Governmental Experts (GGE) on LAWS in November 2017, the AI research company DeepMind released a paper on a new algorithm called AlphaGo Zero that learned to play the Chinese strategy game Go without any human training data.[3] This was a significant technological achievement from the previous iteration of AlphaGo, which required initial training on 30 million moves from human games.[4] AlphaGo Zero demonstrated the ability to learn a complex game from scratch given only the rules of the game. Within three days of self-play, AlphaGo Zero was able to achieve superhuman performance and beat the previous version 100 games to zero.[5] This development was a significant milestone in basic AI research. Although AlphaGo Zero does not have direct military applications, it suggests that current AI technology can be used to solve narrowly defined problems provided that there is a clear goal, the environment is sufficiently constrained, and interactions can be simulated so that computers can learn over time.

Yet even while policymakers and AI scientists are trying to understand how fundamental advances in machine learning are changing the art of the possible for AI, new advances continue. Only a few weeks after the November 2017 GGE, DeepMind released another algorithm in December 2017. This latest version, called AlphaZero, learned to play the strategy games Go, Shogi, and chess all without any human training data. Most significantly, AlphaZero is a single "general purpose" learning algorithm that learned to play all three games. (A separate instance of AlphaZero needed to be trained for each game.)[6] This advancement demonstrates a degree of generality in the method used in AlphaZero—deep reinforcement learning—at least within the domain of strategy games.

States face the daunting task of trying to understand the legal, policy, ethical, strategic, and other considerations of a technology that is rapidly evolving. This paper is intended to be an introductory primer for non-technical audiences on the current state of AI and machine learning, designed to

[1] Within the framework of the Convention on Certain Conventional Weapons (CCW). See https://www.unog.ch/80256EE600585943/(httpPages)/8FA3C2562A60FF81C1257CE600393DF6?OpenDocument.

[2] Electronic Frontier Foundation, "Imagenet Image Recognition," AI Progress Measurements: Vision, https://www.eff.org/ai/metrics#Vision.

[3] David Silver et al., "Mastering the game of Go without human knowledge," *Nature,* vol. 550, pp. 354–359; DeepMind, "AlphaGo Zero: Learning from scratch", no date, https://deepmind.com/blog/alphago-zero-learning-scratch/.

[4] Demis Hassabis, "AlphaGo: Using machine learning to master the ancient game of Go", 27 January 2016, https://blog.google/topics/machine-learning/alphago-machine-learning-game-go/.

[5] DeepMind, "AlphaGo Zero: Learning from scratch", op. cit.

[6] David Silver et al., "Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm", 5 December 2017, https://arxiv.org/pdf/1712.01815.pdf.

support the international discussions on the weaponization of increasingly autonomous technologies. Because of the swiftly changing nature of this field, this paper can only provide a snapshot in time. Many of the underlying concepts about AI and machine learning, however, are likely to remain applicable.

What is artificial intelligence?

Artificial intelligence is the field of study devoted to making machines intelligent.[7] Intelligence measures a system's ability to determine the best course of action to achieve its goals in a wide range of environments.[8] In practice, the field of AI has advanced by AI researchers picking difficult problems, such as playing chess, and then trying to build machines that can accomplish these tasks. Over time, AI researchers have moved from relatively simple games such as tic-tac-toe (noughts and crosses, beaten in 1952), to games of increasing complexity, including checkers (1994), chess (1997), the television game show Jeopardy (2011), some Atari games (2014), Go (2016), poker (2017), and the computer game Dota 2 (2017).[9] AI systems have also begun to perform tasks with many real-world applications, including driving vehicles, image recognition, voice recognition, language translation, and medical diagnoses.

One effect of this ever-advancing field of AI is that at any point in time, only the most cutting edge machines are usually labelled "intelligent". In the 1960s, chess-playing computers were at the forefront of AI research. Today, chess playing machines would not generally be called "intelligent" machines—they are merely computer programs. Similarly, airplane autopilots, tax preparation software, automobile cruise control, smartphones, search engines, and many other software tools have some measure of intelligence. Yet because of their familiarity, they are often referred to as "automation" or simply "computer programs". Often once computers can perform a task, that task is no longer considered to require "intelligence". The machines are merely performing "computation". In this way, the standard for machines being considered "intelligent" is constantly moving. For this reason, it is most useful to think of machines as existing along a spectrum of intelligence without a clear boundary for when a machine should be labelled "intelligent".

Machine learning

Machine learning is one approach to creating intelligent machines. Not all AI systems use machine learning. For example, the AI system Libratus that beat the top human poker players in 2017 does not use learning techniques. However, for many applications, machine learning can be a powerful method for achieving intelligent behaviour.

Rather than follow a proscribed set of *if–then* rules for how to behave in a given situation, learning machines are given a goal to optimize—for example, winning at the game of chess. Learning machines then fine-tune their behaviour based on data in order to optimize the likelihood of achieving their goal. In this way, the machine learns from the data.[10]

---

[7] Adapted from Nils J. Nilsson, *The Quest for Artificial Intelligence: A History of Ideas and Achievements,* Cambridge University Press, 2010.

[8] Adapted from Shane Legg and Marcus Hutter, "A Collection of Definitions of Intelligence", Technical Report IDSIA-07-07, 15 June 2007, p. 9.

[9] Hassabis, "AlphaGo: Using machine learning to master the ancient game of Go", op. cit.; Noam Brown and Tuomas Sandholm, "Superhuman AI for heads-up no-limit poker: Libratus beats top professionals," *Science,* 17 December 2017; OpenAI, "Dota 2", 11 August 2017, https://blog.openai.com/dota-2/.

[10] Adapted from Tom Michael Mitchell, *The Discipline of Machine Learning*, vol. 9, Carnegie Mellon University, School of Computer Science, Machine Learning Department, 2006; Ben Buchanan and Taylor Miller, "Machine Learning for Policymakers: What It Is and Why It Matters",Belfer Center, 2017, https://www.belfercenter.org/sites/default/files/files/publication/MachineLearningforPolicymakers.pdf.

There are many subcategories of machine learning:

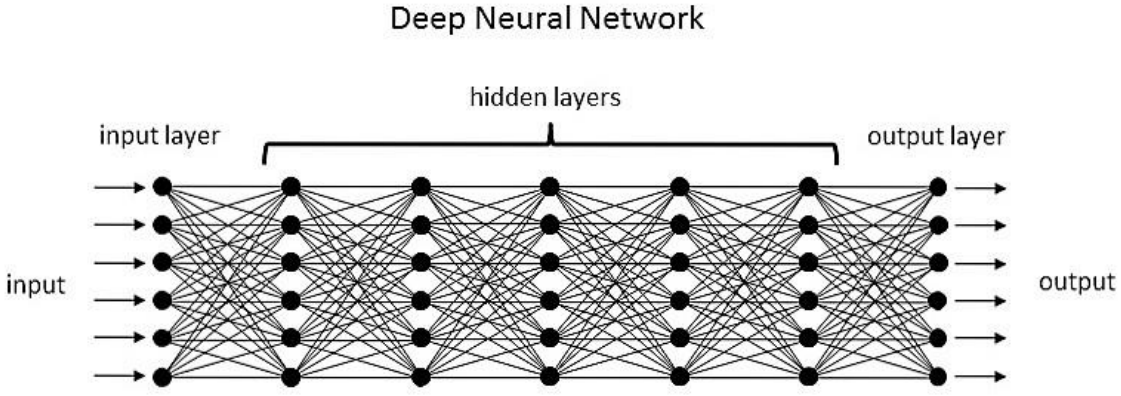
Supervised learning is a machine learning technique that makes use of labelled training data.[11] For example, image classifiers can learn to recognize images by looking at millions of labelled pictures. Over time, these image classifiers can learn to associate subtle patterns in images with certain labels, such as “cat” or “bus”.

Unsupervised learning is a machine learning technique that involves learning from unlabelled data based on the identification of patterns.[12] This technique can be useful if AI researchers do not know what patterns they want to find within data. Based on identifying patterns within the structure of the data, the machine can cluster the data into discrete categories, even if those categories are not labelled. For example, a learning system could analyse financial data and identify different categories of financial transactions. This could be a valuable method for finding anomalous and potentially fraudulent financial transactions, even if the AI researchers do not know what type of financial activity they are seeking to identify. When data shows activity over time, patterns within the data can also be used for predictive analysis, such as weather forecasting.

Reinforcement learning is a machine learning technique in which an agent learns by interacting with its environment.[13] An agent takes an action, observes the effect on its environment, and then determines whether that action helped it achieve its goal. This was the method used by AlphaZero to play chess, Go, and Shogi. AlphaZero played many games against itself and learned over time which moves increased the probability of winning. Reinforcement learning has also been used to train algorithms to play Atari and Nintendo computer games.

Deep learning

Deep learning is a type of machine learning approach that uses deep neural networks. Deep neural networks can be used for supervised, unsupervised, or reinforcement learning. Similarly, machine learning can be done without using neural networks. Deep neural networks have proven to be a powerful tool for learning from large amounts of data, however, and have been used to perform varied tasks such as object recognition, natural language processing, and language translation.



A deep neural network has hidden layers between the input and output layers. Some deep neural networks can have as many as 150 or more hidden layers.

Source: courtesy of Paul Scharre

[11] Adapted from Mehryar Mohri, Afshin Rostamizadeh and Ameet Talwalkar, *Foundations of Machine Learning*, MIT Press, 2012.
[12] Buchanan and Miller, “Machine Learning for Policymakers What It Is and Why It Matters”, op. cit.
[13] Adapted from Richard S. Sutton and Andrew G Barto, *Reinforcement Learning: An Introduction*, vol. 1, MIT Press 1998.

Neural networks rely on a "connectionist" approach loosely inspired by biological neurons. Information flows into an input layer of artificial neurons, which are then connected to other artificial neurons through a network. Signals flow through the network to an output layer. The network "learns" by adjusting the strength of the connections over time in order to optimize the paths through the network to achieve a given output. For example, a neural network used for image classification would take as input the pixels of an image, then output a label for the image, such as "ball" or "lamp". A "deep" neural network is one that has many layers between the input and output layer.

Deep learning requires large amounts of data to train a network. For example, the ImageNet database that is used to train image classifiers has over 14 million labelled images divided into 20,000 categories.[14] The database organizers aim to assemble an average of 1,000 images per category.[15] When large amounts of data are not available, AI researchers can often train neural networks using "synthetic data" created via computer simulations. AlphaGo Zero did not use any initial training data from human games of Go and instead learned to play through 4.9 million games of self-play.[16] Because large numbers of games of Go could be simulated through self-play, this was a viable approach for training the network. For problems that lack large amounts of data or the ability to generate synthetic data through simulations, however, these methods may not be as effective.

Computing resources

One interesting feature of machine learning is that, once trained, the computing power required to use a trained algorithm is often significantly less than the computing power required to train the algorithm in the first place. For example, AlphaZero used 5,000 Tensor Processing Units (TPU) to generate self-play games and another 64 TPUs to train the neural network. Once trained, however, AlphaZero required only 4 TPUs to play games.[17] (A TPU is a type of application-specific integrated circuit (ASIC)—a computer chip designed for a specific function—that is optimized for machine learning with neural networks.)

It can often take significant amounts of computing power and data to train a neural network. Once trained, however, the network can be used with vastly less computing power and without the initial training data. This means that while the data and computing power needed to create the most powerful new AI algorithms may be limited to a smaller set of actors, a larger number of groups and individuals may often be able to use already-trained algorithms. The AI research community is extremely open and there are many open-source libraries available online from which researchers can download trained neural networks and use them for a variety of applications.[18]

Other AI and machine learning methods

While many recent gains in AI have come from deep learning—and deep reinforcement learning in particular—there are many other approaches to machine learning and AI.[19] A full survey of all of the

[14] ImageNet, "About ImageNet: Summary and Statistics", http://image-net.org/about-stats.
[15] ImageNet, "About ImageNet: Overview", http://image-net.org/about-overview.
[16] Silver et al., "Mastering the game of Go without human knowledge", op. cit., p. 355.
[17] Silver et al., "Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm", op. cit., p. 4.
[18] For some examples, see: "TensorFlow", https://www.tensorflow.org/; "Caffe", http://caffe.berkeleyvision.org/; "Neural Network Libraries by Sony", https://nnabla.org/; and "Apache MXNet", http://mxnet.incubator.apache.org/index.html.
[19] Kai Arulkumaran et al., "A Brief Survey of Deep Reinforcement Learning", *IEEE Signal Processing Magazine*, 28 September 2017, https://arxiv.org/pdf/1708.05866.pdf.

many AI methods is beyond the scope of this paper.[20] However, a handful of different methodological approaches to AI are listed below as examples:

- Evolutionary or genetic algorithms;
- Inductive reasoning;
- Computational game theory;
- Bayesian statistics;
- Fuzzy logic;
- Hand-coded expert knowledge; and
- Analogical reasoning.

Over time, the popularity of these and other methods in the AI research community has waxed and waned. Neural networks were first introduced in the 1960s, but have become more viable in recent years in part because of the availability of large data sets and computing power to train neural networks. Algorithms for using deep neural networks have also improved over time, and both computing power and better algorithms are important sources of progress in AI.[21] As the field of AI progresses, other techniques or combinations of techniques are sure to be invented.

Increasing autonomy

Autonomy is, like intelligence, another attribute of both people and machines. Intelligence is a system's ability to *determine the best course of action* to achieve its goals. Autonomy is the *freedom* a system has in accomplishing its goals. Greater autonomy means more freedom, either in the form of undertaking more tasks, with less supervision, for longer periods in space and time, or in more complex environments. Intelligence and autonomy are different attributes of machines (and people). A highly intelligent system could have little autonomy. A chess grandmaster whose hands are tied may know the best course of action to win a game, but lack the freedom to make a move.

Intelligence is related to autonomy in that more intelligent systems are capable of deciding the best course of action for more difficult tasks in more complex environments. This means that more intelligent systems *could* be granted more autonomy and would be capable of successfully accomplishing their goals. Intelligent systems do not spontaneously become more autonomous, though. Humans choose to give machines more autonomy as they become more intelligent. For example, as self-driving cars become more capable, they will become trusted to operate with less human supervision over time.

Narrow versus general AI

All AI systems in existence today fall under the broad category of "narrow AI". This means that their intelligence is limited to a single task or domain of knowledge. A chess-playing program cannot play checkers, Go, or other similar strategy games. (While AlphaZero used a single general-purpose algorithm to learn to play chess, Go, and Shogi, different versions of AlphaZero had to be trained for

[20] For more on different approaches to artificial intelligence and machine learning, see: Buchanan and Miller, "Machine Learning for Policymakers What It Is and Why It Matters", op. cit.; and Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World,* Basic Books, 2015.

[21] As a specific example, DeepMind cited algorithmic progress as an important source of improvement in its iterations of AlphaGo over time: DeepMind, "AlphaGo Zero: Learning from scratch", op. cit. For a more general analysis of algorithmic vs. hardware improvements as drivers of AI progress, see Katja Grace, "Algorithmic Progress in Six Domains", Machine Intelligence Research Institute, https://intelligence.org/files/AlgorithmicProgress.pdf.

each game.[22]) Reportedly, AlphaGo even performs poorly when playing Go on a different sized board.[23] A key technical reason for this limitation is the problem of "catastrophic forgetting".[24] When deep neural networks attempt to learn a new task, they generally "forget" the ability to perform the old task. This prevents the machine from continually learning new tasks. A related problem is that AI systems generally are unable to transfer skills learned for one task to another, related task. While AI researchers are working on this problem, to-date it remains a major limitation for AI systems.[25]

The effect of this limitation is that while many AI systems can achieve superhuman performance in their narrow domains, general-purpose learning systems that can acquire knowledge in multiple domains remain an elusive goal of AI researchers. This goal is sometimes stated in various ways, with researchers using terms such as artificial general intelligence (AGI), general-purpose AI, cross-domain AI, Third Wave AI, transformative AI, high-level machine intelligence, or true AI. Depending on the context in which they are used, these terms often mean slightly different things, but all envision a future form of AI that is able to overcome the limitations of today's narrow, task-specific AI. These goals are often stated in human-centric terms, such as "human-level AI" that can equal or surpass humans in all activities or domains of knowledge. There is no reason to believe that advanced AI systems would necessarily think like humans or acquire knowledge in the same domains as humans, though. For example, a hypothetical future general-purpose learning system could achieve superhuman performance in some tasks across multiple domains, but lag behind human performance in other areas depending on which tasks were most important to achieving its goals. In fact, experience with AI systems to-date shows that they often think in ways that are strange and counterintuitive relative to human cognition.[26]

Superintelligence

Superintelligence is a term used to refer to a hypothetical future AI system that could vastly exceed human intelligence in all possible domains. Such a system would be as dominant over humans as AI systems are today in chess or Go, but in all domains of knowledge and expertise. Some scientists have warned of the dangers of superintelligent AI.[27] Physicist Stephen Hawking has said, "The

[22] Silver et al., "Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm", op. cit., p. 4.

[23] Bob van den Hoek, 13 May 2016, answer on the question, "What would happen if DeepMind's AlphaGo had to play a match of GO (same rules), but on a different size of square board (say 20x20 instead of 19x19)", *Quora*, https://www.quora.com/What-would-happen-if-DeepMinds-AlphaGo-had-to-play-a-match-of-GO-same-rules-but-on-a-different-size-of-square-board-say-20x20-instead-of-19x19.

[24] Ronald Kemker et al., "Measuring Catastrophic Forgetting in Neural Networks", 9 November 2017, https://arxiv.org/pdf/1708.02072.pdf; James Kirkpatrick, "Overcoming catastrophic forgetting in neural networks", 25 January 2017, https://arxiv.org/pdf/1612.00796.pdf.

[25] For some examples of recent progress, see Melvin Johnson et al., "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation", 21 August 2017, https://arxiv.org/pdf/1611.04558.pdf; and Lasse Espeholt et al., "IMPALA: Scalable Distributed Deep-RL with Importance Weighted Actor-Learner Architectures", 9 February 2018, https://arxiv.org/pdf/1802.01561.pdf. For a simpler explanation of these technical papers, see Mike Schuster, "Zero-Shot Translation With Google's Multi-Lingual Neural Machine Translation System", Google Research Blog, 22 November 2016, https://research.googleblog.com/2016/11/zero-shot-translation-with-googles.html; and DeepMind, "IMPALA: Scaleable Distributed DeepRL in DMLab-30", 5 February 2018, https://deepmind.com/blog/impala-scalable-distributed-deeprl-dmlab-30/.

[26] David Berreby, "Artificial Intelligence is Already Weirdly Inhuman", *Nautilus*, 6 August 2015, http://nautil.us/issue/27/dark-matter/artificial-intelligence-is-already-weirdly-inhuman.

[27] For more on the potential dangers of superintelligence, see Eliezer Yudkowsky, "Artificial Intelligence as a Positive and Negative Factor in Global Risk", 2007, http://yudkowsky.net/singularity/ai-risk; Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies,* Oxford University Press, 2014; Stuart Russell, "Of Myths and Moonshine", *Edge,* 14 November 2014, https://www.edge.org/conversation/the-myth-of-ai#26015; Scott Alexander, "No Time Like the Present for AI Safety Work", Slate Star Codex, 29 May 2015, http://slatestarcodex.com/2015/05/29/no-time-like-the-present-for-ai-safety-work/; Scott Alexander, "AI Researchers on AI Risk", Slate Star Codex, 22 May 2015, http://slatestarcodex.com/2015/05/22/ai-researchers-on-ai-risk/; Robin Hanson, "Foom Justifies AI Risk Efforts Now"; Overcoming Bias blog on overcomingbias.com, 3 August 2017, http://www.overcomingbias.com/2017/08/foom-justifies-ai-risk-efforts-now.html.

development of full AI could spell the end of the human race."[28] Microsoft co-founder Bill Gates has said he believes AI "will usher in growth and productivity in the near term, but has long-term risks" and that he is "concerned about superintelligence".[29]

Given the limitations of AI systems today, there appears to be no danger of superintelligent AI in the near-term. Some researchers have hypothesized that a future machine that has human-level abilities across many domains, including in building new AI systems, could build a slightly more intelligent machine, which could in turn build an even more intelligent machine, and so on. This would be analogous to the self-improvement that existing AI systems have used to surpass human abilities in narrow domains such as chess and Go, although it should be noted that current systems are still designed and built by humans. Whether superintelligence is possible and on what timescale is unclear. Surveys of AI researchers suggest that the majority of AI researchers believe human-level AI is theoretically possible.[30] A number of recent surveys predict human-level AI around the 2040s, and the aggregate forecast among AI researchers is that human-level AI has a 50% probability of being developed in the next 45 years. Estimates among AI researchers vary widely on the timescales at which human-level AI might be achieved, however, with some believing it is possible in the next decade and others believing it is unlikely within the next century.[31] Similarly, if human-level AI *could* be developed, AI researchers have varying views on whether superintelligence would follow quickly or slowly afterward.[32] Given the tremendous uncertainty in what it would take to create more advanced forms of AI, combined with the lack of scientific knowledge about how human intelligence works, these predictions should be taken with a healthy amount of scepticism.

Practical uses of AI today

Artificial intelligence is an enabling technology that is applicable in a wide range of areas, much like electricity or the internal combustion engine.[33] During the industrial revolution, these technologies enabled the creation of purpose-built machines that were stronger than humans for certain tasks. The AI revolution is enabling the creation of purpose-built machines that outperform humans at certain tasks.

While the long-term trajectory of AI is uncertain, in the near-term there are many applications for existing narrow AI. Around the world, researchers are applying progress in AI to build machines to help solve a variety of difficult problems.[34]

---

[28] Rory Cellan-Jones, "Stephen Hawking Warns Artificial Intelligence Could End Mankind", *BBC News*, 2 December 2014, sec. Technology, http://www.bbc.com/news/technology-30290540.

[29] Peter Holley, "Bill Gates on Dangers of Artificial Intelligence: 'I Don't Understand Why Some People Are Not Concerned'", *Washington Post,* 29 January 2015, https://www.washingtonpost.com/news/the-switch/wp/2015/01/28/bill-gates-on-dangers-of-artificial-intelligence-dont-understand-why-some-people-are-not-concerned/?utm_term=.de73c28151ca.

[30] Vincent Müller, Nick Bostrom, "Future Progress in Artificial Intelligence: A Survey of Expert Opinion", Future of Humanity Institute, 2014, p. 11.

[31] Katja Grace et al., "When Will AI Exceed Human Performance? Evidence From AI Experts", 30 May 2017, https://arxiv.org/pdf/1705.08807.pdf; Katja Grace, "Update on all the AI Predictions", AI Impacts, 5 June 2015, https://aiimpacts.org/update-on-all-the-ai-predictions.

[32] Ben Goertzel, "The Hard Takeoff Hypothesis", The Multiverse According to Ben, 13 January 2011, http://multiverseaccordingtoben.blogspot.co.uk/2011/01/hard-takeoff-hypothesis.html; Eliezer Yudkowsky, "Hard Takeoff," LessWrong, 2 December 2008, https://wiki.lesswrong.com/wiki/AI_takeoff; Robin Hanson and Eliezer Yudkowsky, "The Hanson-Yudkowsky AI–Foom Debate," http://intelligence.org/files/AIFoomDebate.pdf; Luke Muelhauser and Anna Salamon, "Intelligence Explosion: Evidence and Import", Machine Intelligence Research Institute, 2012, http://intelligence.org/files/IE-EI.pdf; Bostrom, *Superintelligence,* op. cit., pp. 95–101.

[33] Marc Benioff, "Marc Benioff: We're on the cusp of an AI revolution", *World Economic Forum*, 15 September 2016, https://www.weforum.org/agenda/2016/09/marc-benioff-were-on-the-cusp-of-an-ai-revolution.

[34] A number of thinkers have suggested AI has the potential to yield disruptive change on the order of another industrial revolution. See, for example, Klaus Schwab, "The Fourth Industrial Revolution: What it Means, How to Respond", *World Economic Forum,* 14 January 2016, https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/; Bank of America – Merrill Lynch, "Robot Revolution – Global Robot & AI Primer", https://www.bofaml.com/content/dam/boamlimages/documents/PDFs/robotics_and_ai_condensed_primer.pdf.

AI systems can be used for a range of purposes, such as:

- Data analytics, such as medical diagnoses;
- Controlling autonomous systems, such as self-driving cars;
- Predicting future trends or behaviour from data;
- Object classification and recognition;
- Detecting anomalous activity, such as in financial transactions;
- Optimizing systems to achieve a goal; and
- Performing simple automated tasks at scale.

In practice, AI systems are often used in conjunction with human decision-making because of the limitations of current AI systems. For example, current top-of-the-line automobiles have a number of automated and partially autonomous features, such as intelligent cruise control, self-parking, and automatic collision avoidance. Even in automobiles with the most advanced autopilots on the market today, humans are still needed for some driving tasks. Over time, this will likely change as machine perception and decision-making improve in self-driving cars. Eventually, it will be possible for cars to drive fully autonomously without any human intervention in a wide variety of environments. Humans will still make higher-level decisions such as the trip destination, however.

Driving is a task that largely hinges on perception, quick decision-making, rapid reaction, and physical skill. Other tasks that require weighing competing values, applying judgment, or interacting with humans are likely to require human involvement in some capacity for much longer periods of time. And even if it is no longer "required" due to technological developments, we may decide it is desirable to maintain human involvement due to reasons rooted, for instance, in psychology, values or ethics. For example, while AI systems help doctors make more accurate medical diagnoses, they are unlikely to replace human doctors entirely. Many tasks that doctors perform, such as helping patients understand their diagnosis, and weighing the risks and benefits of potential treatment options, are much harder tasks for machine intelligence today as they require empathy and an appreciation for context.

Conclusion

Recent gains in AI enable machines that are sophisticated and can solve many problems, in some cases measurably better than humans.[35] Nevertheless, AI systems still have significant limitations and vulnerabilities. At present, AI systems lack the ability to engage in general-purpose reasoning or transfer skills learned in one task to another, related task. This makes their intelligence "brittle". Current domain-specific AI systems exhibit narrow intelligence. In practice, this means that it remains the responsibility of humans to know when and where to employ these systems, as well as their limitations.

Additionally, many advance AI and machine learning methods suffer from problems of predictability, explainability, verifiability, and reliability. Machine learning systems that are given the wrong goals or fed incorrect or biased data can learn the wrong behaviours, sometimes in surprising ways.[36] These are important vulnerabilities in existing AI methods that must be accounted for as these systems are increasingly used in the real world. As AI systems transition from

[35] For specific examples, see Electronic Frontier Foundation, "AI Progress Measurement", https://www.eff.org/ai/metrics.

[36] For more on these and other safety problems in current AI systems, see, Dario Amodei et al., "Concrete Problems in AI Safety", 25 July 2016, https://arxiv.org/pdf/1606.06565.pdf.

research laboratories to real-world environments, verifying safe behaviour will be increasingly important.

Over time, as AI systems become more sophisticated, they will be incorporated into a variety of applications across human society. Even when systems perform correctly, their use will often raise difficult legal and ethical questions that society must address. Some of these issues may stem from the consequences of transferring control of tasks that were previously done by humans to machines, and the implications for accountability and responsibility. Other issues revolve around the necessity to codify rules for machine behaviour that may have been implicit in the past or subject to individual human judgment. For example, should a self-driving car be permitted to violate the speed limit if traffic conditions are such that driving with the flow of traffic (above the speed limit) is safer?

These issues take on added weight when considering the use of AI-enabled systems in military contexts. War is an adversarial environment where adaptation, flexibility, and surprise are common. These pose unique challenges to verifying and testing AI-enabled systems.[37] Additionally, the consequences for failure in war can be high. Accidents can have tragic consequences when lethal military systems are involved.[38] At the same time, sometimes the application of lethal force (or being prepared to use force) is required to prevent greater harm.

These challenges point to the continued need for an ongoing, robust discussion among States on the weaponization of increasingly autonomous technologies. Because of the complex legal, moral, ethical, and other issues raised by AI systems,[39] policymakers are best served by a cross-disciplinary dialogue that includes scientists, engineers, military professionals, lawyers, ethicists, academics, members of civil society, and other voices. Including diverse perspectives in ongoing discussions surrounding lethal autonomous weapons can help ensure that militaries use emerging technologies in responsible ways.

[37] For examples of how AI could be intentionally manipulated in adversarial environments, see Future of Humanity Institute et al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation", February 2018. https://maliciousaireport.godaddysites.com/.

[38] See UNIDIR, *Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies*, Observation Report no. 5, 2016; Paul Scharre, *Autonomous Weapons and Operational Risk,* Center for a New American Security, 2016.

[39] Some of these issues as they relate to the application of AI in military systems were raised in the "Food-For-Thought Paper" submitted by the Chairperson of the 2017 GGE on LAWS, 4 September 2017, UN document CCW/GGE.1/2017/WP.1, http://undocs.org/ccw/gge.1/2017/WP.1.

Appendix: Recommended resources

For further reading on artificial intelligence and machine learning, see:

Amodei, Dario et al., “Concrete Problems in AI Safety”, 25 July 2016, https://arxiv.org/pdf/1606.06565.pdf.

Buchanan, Ben and Miller, Taylor, “Machine Learning for Policymakers: What it is and Why it Matters”, Harvard Belfer Center, June 2017. https://www.belfercenter.org/sites/default/files/files/publication/MachineLearningforPolicymakers.pdf.

Domingos, Pedro, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, Basic Books, 2015.

Electronic Frontier Foundation, “AI Progress Measurement”, https://www.eff.org/ai/metrics.

Future of Humanity Institute et al., “The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation”, February 2018. https://maliciousaireport.godaddysites.com/.

Scharre, Paul, *Army of None: Autonomous Weapons and the Future of War,* W.W. Norton & Company, 2018.

Stone, Peter et al., "Artificial Intelligence and Life in 2030", *One Hundred Year Study on Artificial Intelligence: Report of the 2015–2016 Study Panel*, Stanford University, September 2016. http://ai100.stanford.edu/2016-report.

UNIDIR Observation Reports on the Weaponization of Increasingly Autonomous Technologies

Framing Discussions on the Weaponization of Increasingly Autonomous Technologies, Observation Report no. 1 (2014).

The Weaponization of Increasingly Autonomous Technologies: Considering how Meaningful Human Control might move the discussion forward, Observation Report no. 2 (2014).

The Weaponization of Increasingly Autonomous Technologies: Considering Ethics and Social Values, Observation Report no. 3 (2015).

The Weaponization of Increasingly Autonomous Technologies in the Maritime Environment: Testing the Waters, Observation Report no. 4 (2015).

Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies, Observation Report no. 5 (2016).

The Weaponization of Increasingly Autonomous Technologies: Concerns, Characteristics and Definitional Approaches, Observation Report no. 6 (2017).

The Weaponization of Increasingly Autonomous Technologies: Autonomous Weapon Systems and Cyber Operations, Observation Report no. 7 (2017).

The Weaponization of Increasingly Autonomous Technologies: Artificial Intelligence, Observation Report no. 8 (2018).

Available for download at bit.ly/UNIDIR_Autonomy

UNIDIR

The Weaponization of Increasingly Autonomous Technologies: Artificial Intelligence

a primer for CCW delegates

The rapidly advancing field of AI and machine learning has significant implications for the role of autonomy in weapon systems. States face the daunting task of trying to understand the legal, policy, ethical, strategic, and other considerations of a technology that is rapidly evolving. This paper is an introductory primer for non-technical audiences on the current state of AI and machine learning, designed to support the international discussions on the weaponization of increasingly autonomous technologies.

No. 8

UNIDIR RESOURCES