sipri

MAPPING THE DEVELOPMENT OF AUTONOMY IN WEAPON SYSTEMS

A primer on autonomy

VINCENT BOULANIN

WORKING PAPER
December 2016

Mapping the development of autonomy in weapon systems

A primer on autonomy

VINCENT BOULANIN

STOCKHOLM INTERNATIONAL
PEACE RESEARCH INSTITUTE

STOCKHOLM INTERNATIONAL PEACE RESEARCH INSTITUTE

SIPRI is an independent international institute dedicated to research into conflict, armaments, arms control and disarmament. Established in 1966, SIPRI provides data, analysis and recommendations, based on open sources, to policymakers, researchers, media and the interested public.

The Governing Board is not responsible for the views expressed in the publications of the Institute.

STOCKHOLM INTERNATIONAL PEACE RESEARCH INSTITUTE

Signalistgatan 9
SE-169 72 Solna, Sweden
Telephone: +46 8 655 97 00
Email: sipri@sipri.org
Internet: www.sipri.org

Acknowledgements

The SIPRI project on Autonomy in Weapon Systems is supported by the Federal Foreign Office of Germany, the Ministry of Foreign Affairs of the Netherlands, the Ministry for Foreign Affairs of Sweden, and the Federal Department for Foreign Affairs of Switzerland.

The author is indebted to the experts who accepted to participate in background interviews for their rich and valuable input. The author would also like to acknowledge Maaike Verbruggen's invaluable contribution to the literature review and data collection. Lastly, the author is grateful to the reviewers for their very constructive feedback.

Responsibility for the information set out in this working paper lies entirely with the author.

Contents

List of boxes, tables and figures

Abbreviations

CCW	Convention on Certain Conventional Weapons
GPS	Global Positioning System
GSM	the Global System for Mobile
HRW	Human Rights Watch
IHL	International humanitarian law
LAWS	Lethal autonomous weapon systems
NASA	National Aeronautics and Space Administration
NGO	Non-governmental organization
UAV	Unmanned aerial vehicles

Preface

Since 2013 the governance of lethal autonomous weapon systems (LAWS) has been discussed internationally under the framework of the 1980 United Nations Convention on Certain Conventional Weapons (CCW), which regulates weapons that may be deemed to have an excessively injurious or indiscriminate effect.[1] Issues of concern include the moral acceptability of LAWS, their potentially negative impact on interstate relations and stability, their possible facilitation of recourse to the use of force, and their compatibility with international humanitarian law and international human rights law.[2]

The discussion is still at an early stage. The question of whether the states parties to the CCW should take formal action on LAWS is not yet officially on the agenda, albeit it is a key part of the ongoing debate. The Campaign to Stop Killer Robots, a coalition of non-governmental organizations (NGOs), and a handful of states are already advocating the adoption of a pre-emptive ban on the development, production and use of LAWS.[3] Most states, however, have expressed that they are not yet ready to discuss this possibility as they are still in the process of understanding the full implications of increasing autonomy in weapon systems.

To support states in their reflection on this issue, and contribute to more concrete and structured discussion on LAWS at the various meetings associated with the CCW, in February 2016 the Stockholm International Peace Research Institute (SIPRI) launched a research project assessing the development of autonomy in military systems in general and weapon systems in particular. The project entitled 'Mapping the development of autonomy in weapon systems' was designed based on the assumption that efforts to develop concepts and practical measures for monitoring and controlling LAWS will remain premature without a better understanding of (*a*) the technological foundations of autonomy, (*b*) the current applications and autonomy capabilities in existing weapon systems, and (*c*) the technological, socio-economical, operational and political factors that are currently enabling or limiting advances in the sphere of LAWS. The project's aim, in that regard, is to provide CCW delegates and the interested public with a 'reality check' on autonomy through a mapping exercise that will answer a series of fundamental questions:

1. What is autonomy? How does it work? How is it created?
2. What are the underlying technologies and where are they available or being developed?
3. What types of autonomous applications are found in existing and forthcoming weapon systems?
4. What are the capabilities of weapons that include some level of autonomy in the target cycle? How are they used or intended to be used, and what are the principles or rules that govern their use?
5. What are the trends that fuel or limit the advance of autonomy in weapon systems?

These questions will be addressed in a series of four working papers. Each of these papers will map the development of autonomy in weapon systems from a different perspective. The first working paper is intended to serve as a primer on the technological foundation of autonomy (technical perspective). The second working paper will

[1] 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, or 'Inhumane Weapons' Convention), with Protocols I, II and III, opened for signature 10 Apr. 1981, entered into force 2 Dec. 1983, <http://treaties.un.org/Pages/CTCTreaties.aspx?id=26>.

[2] Docherty, B., *Losing Humanity: The Case Against Killer Robots* (Human Rights Watch/International Human Rights Clinic: Washington, DC, 2012); and Sharkey, N., 'Saying "no!" to lethal autonomous targeting', *Journal of Military Ethics*, vol. 9, no. 4 (2010).

[3] Bolivia, Cuba, Ecuador, Egypt, Ghana, the Holy See, Pakistan, Palestine and Zimbabwe expressed clear support for a ban on LAWS. Croatia, Ireland and Sri Lanka are open to considering a ban.

map out the innovation ecosystem that is driving the advance of autonomy in weapon systems (economic perspective). The third working paper will provide a systematic assessment of the capabilities of autonomy in existing weapon systems (operational perspective), while the final working paper will discuss the political drivers and obstacles to the adoption of autonomy in weapon systems (political perspective). These four papers will be integrated into a final report to be published in early 2017.

1. Introduction

Since 2013 the governance of lethal autonomous weapon systems (LAWS) has been discussed under the framework of the 1980 United Nations Convention on Certain Conventional Weapons (CCW).[1] The discussion is still at an early stage, with most states parties still in the process of understanding the issues at stake—beginning with the fundamental questions of what constitutes 'autonomy' and to what extent it is a matter of concern in the context of weapon systems and the use of force. States parties that took the floor during the three informal meetings of experts that were organized in 2014, 2015 and 2016 under the auspices of the CCW presented rather different interpretations of the defining characteristics of autonomy, thereby contributing to confusion as to what types of systems and legal, ethical, operational and security concerns were actually up for discussion. Some states define 'autonomy' in a way that encompasses a number of existing systems. Others define the term more narrowly, which excludes current systems and can be applied only to systems that are not as yet in existence. Thus, a number of states parties have stressed that future discussions could usefully benefit from further investigation into the conceptual and technical foundations of the meaning of 'autonomy'.

This working paper is an attempt to respond to that demand. It aims to clarify some basic understandings about autonomy: what it is, how it applies to weapon systems, how it works, how it is created and what the key technological enablers are. It is based on a substantial review of the literature as well as a background series of interviews with experts from various expert communities. The next section (section 2) reviews existing interpretations of the concept of autonomy. Section 3 maps out possible applications of autonomy in weapon systems, while section 4 identifies the underlying capabilities and technologies on which autonomy may be created. Section 5 discusses the current state of autonomy. The concluding section (section 6) presents some takeaway points for future discussions on LAWS within the framework of the CCW.

[1] 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, or 'Inhumane Weapons' Convention), with Protocols I, II and III, opened for signature 10 Apr. 1981, entered into force 2 Dec. 1983, http://treaties.un.org/Pages/CTCTreaties.aspx?id=26.

2. Searching for definition: what is autonomy?

I. Autonomy: a three-dimensional concept

In simple terms 'autonomy' can be defined as the ability of a machine to execute a task, or tasks, without human input, using interaction of computer programming with the environment.[1] 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.

However, autonomy is a relative notion: within and across relevant disciplines, be it engineering, robotics or computer science, experts have a different understanding of when a system or a system's function may or may not be deemed autonomous. As previously identified by Scharre, these approaches can be sorted into three categories.[2]

The human-machine command-and-control relationship

A very common approach for assessing autonomy relates to the extent to which humans are involved in the execution of the task carried out by the machine. With this approach the systems can be classified into three categories. Systems that require human input at some stage of the task execution can be referred to as 'semi-autonomous' or 'human in the loop'. Systems that can operate independently but are under the oversight of a human who can intervene if something goes wrong (e.g. malfunction or systems failure) are called 'human-supervised autonomous' or 'human on the loop'. Machines that operate completely on their own and where humans are not in a position to intervene are usually referred to as 'fully autonomous' or 'human out of the loop'. The concept of 'sliding autonomy' is sometimes also employed to refer to systems that can go back and forth between semi-autonomy and full autonomy, depending on the complexity of the mission, external operating environments and, most importantly, legal and policy constraints.

The sophistication of the machine's decision-making process

A more technical approach to autonomy relates to the actual ability of a system to exercise control over its own behaviour (self-governance) and deal with uncertainties in its operating environment.[3] From this standpoint, systems are often sorted into three major categories: automatic, automated and autonomous systems. 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 (e.g. robotic arms used in the manufacturing industry). Machines that can cope with variations in their environment and exercise control over their actions can either be described as automated or autonomous. What distinguishes an automated system from an autonomous system is a contentious issue. Some experts see the difference in terms of degree of self-governance, and view autonomous systems merely as more complex and intelligent forms of automated systems. Others see value in making a clear distinction between the two concepts. Williams,

[1] This definition is based on one previously proposed by Andrew Williams. Williams, A., 'Defining autonomy in systems: challenges and solutions', eds A. Williams and P. Scharre, *Autonomous Systems: Issues for Defence Policymakers* (NATO Headquarters Allied Command: Norfolk, VA, 2015).

[2] Scharre, P., 'The opportunity and challenge of autonomous systems', eds Williams and Scharre (note 1), p. 56.

[3] E.g. 'Autonomy refers to a robot's ability to accommodate variations in its environment. Different robots exhibit different degrees of autonomy; the degree of autonomy is often measured by relating the degree at which the environment can be varied to the mean time between failures, and other factors indicative of robot performance.' Thrun, S., 'Toward a framework for human-robot interaction', *Human-Computer Interaction*, vol. 19, no. 1–2 (2004), pp. 9–24.

for instance, presents an 'automated system' as a system that 'is programmed to logically follow 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'. On the other hand, an 'autonomous system':

> is capable of understanding higher-level intent and direction. From this understanding and its perception of its environment, such a system can take appropriate action to bring about a desired state. It is capable of deciding a course of action, from a number of alternatives, without depending on human oversight and control, although these may still be present. Although the overall activity of an autonomous unmanned aircraft will be predictable, individual actions may not be.[4]

While the distinction between automatic, automated and autonomous can be conceptually useful, in practice it is has proved difficult to measure and therefore determine whether a system falls within one of the three categories. Moreover, the definitions of and boundaries between these three categories are contested within and between the expert communities.

The types of decisions or functions being made autonomous

A third dimension to consider focuses on the types of decisions or functions that are made autonomous within a system. This 'functional' approach is not incompatible with the two other approaches; it acknowledges simply that referring to autonomy as a general attribute of systems is imprecise, if not meaningless, as it is the nature of the tasks that are completed autonomously by a machine that primarily matters, not the level of autonomy of the systems as a whole. Autonomy is best understood in relation to the types of tasks that are executed at the subsystems/function level.[5] Some functions in weapon systems may be made autonomous without presenting significant ethical, legal or strategic risks (e.g. navigation), while others may be a source of greater concern (e.g. targeting).[6]

Autonomy in weapon systems: a situated approach

This working paper and those that follow favour a 'functional approach' to autonomy. The notable merit of this approach is that it enables a flexible examination of the challenges posed by autonomy in weapon systems. It recognizes that the human-machine command-and-control relationship and the sophistication of a machine's decision-making capability may vary from one function to another. Some functions may require a greater level of self-governance than others, while human control may be exerted on some functions but not others depending on the mission complexity and the external operating environment as well as regulatory constraints. Also, the extent of human operators' control or cancel functions may change during the system's mission.

Thus, it could be said that the focus of the research presented in the working papers is on the development of autonomy *in* weapon systems rather than the development of autonomous systems per se. The ambition is to discuss the development and application of autonomy in a large range of weapon systems in general, not just the few

[4] Mindell, D., *Our Robots, Ourselves, Robotics and the Myths of Autonomy* (Viking: New York, NY, 2015), p. 12.

[5] United Nations Institute for Disarmament Research (UNIDIR), *Framing Discussions on the Weaponization of Increasingly Autonomous Technologies*, UNIDIR Resources no. 1 (UNIDIR: Geneva, 2014).

[6] NATO, *Uninhabited Military Vehicles: Human Factors Issues in Augmenting the Force*, NATO Technical Report RTO-TR-HFM-078 (NATO: 2007); Vignard, K., 'Statement by the United Nations Institute for Disarmament Research', 2016 CCW Informal Meeting of Experts on Lethal Autonomous Weapon Systems, Geneva, 12 Apr. 2016, <http://www.unog.ch/80256EDD006B8954/(httpAssets)/86C96CC8C7A932DCC1257F930057C0E3/$file/2016_LAWS+MX_GeneralExchange_Statements_UNIDIR.pdf>; and Gillespie, A., 'Humanity and lethal robots: an engineering perspective', ed. G. Verdirame et al., *SNT Really Makes Reality, Technological Innovation, Non-Obvious Warfare and the Challenges of International Law* (King's College London: London, Forthcoming).

types of weapon systems that may be classified as autonomous according to some existing definitions (current definitions of autonomous weapon systems are presented in Appendix A; the types of weapon systems that are sometimes described as autonomous are presented in section 3).[7]

[7] For a number of experts, the term 'autonomous weapon systems' is actually a misnomer. Stensson and Jansson argue for instance that the concept of 'autonomy' is maladaptive as it implies, philosophically, qualities that technologies cannot have. For them, machines, by definition, cannot be autonomous. Stensson, P. and Jansson, A., 'Autonomous technology: source of confusion: a model for explanation and prediction of conceptual shifts', *Ergonomics*, vol. 57, no 3 (2014), pp. 455–70. The concept of autonomous systems has also caused complex and contentious debate regarding the level at which a system may be deemed truly autonomous. In a report dated 2012 the US Department of Defense's Defense Science Board concluded that defining levels of autonomy was a waste of time and money, and tended to reinforce fears of unbounded autonomy. The report noted that discussion of levels of autonomy 'deflects focus from the fact that all autonomous systems are joint human-machine cognitive systems … all systems are supervised by humans to some degree … There are no fully autonomous weapons systems as there are no fully autonomous sailors, airmen, or marines'. US Department of Defense (DOD), Defense Science Board, *Task Force Report: Role of Autonomy in DoD Systems* (DOD: Washington, DC, 2012), pp. 23–24.

3. Situating autonomy: where is autonomy in weapon systems?

I. Applications of autonomy in weapon systems

Autonomy is a characteristic that can be attached to a large variety of functions in weapon systems. These may be sorted into five generic task areas: (*a*) mobility, (*b*) health management, (*c*) interoperability, (*d*) battlefield intelligence, and (*e*) use of force (see table 1).

Mobility includes various types of functions that allow systems to govern and direct their own motion within their operating environment. Key applications of autonomy for mobility include navigation, take-off and landing, obstacle avoidance, and return to base in case of loss of communication.

The second area, *health management*, regroups functions that allow systems to manage their functioning or survival. One example is power management: when a system detects that its power resources are low, it can engage and manage the process of recharging or refuelling completely independently. Other possible applications include autonomous fault detection and self-repair.

The third area, *interoperability* (focusing here on machine autonomy) is concerned with the ability of a machine to execute a task in collaboration with other machines or humans. Swarming is one notable example of machine-to-machine collaboration consisting of making large numbers of simple or low-cost physical robots execute a task in concert, which can be done in a centralized or decentralized way.[1] A number of experts foresee that developments in swarming will have a fundamental impact on future warfare, as it would enable the application of force with greater coordination, intelligence, mass and speed.[2] In terms of human-machine collaboration, one key concrete application of autonomy is to enable the use of natural language (either speech or gesture) for command and control. Voice command and control is already in use in some weapon platforms, but so far it is limited to the execution of non-critical tasks.

The fourth area, *battlefield intelligence*, refers to on-board functions that allow weapon systems to find and analyse data of tactical or strategic relevance on the battlefield. The data may then serve to guide decision making by either the operators or military command.

The fifth and most critical category, *use of force*, refers specifically to functions that enable weapon systems to search for, detect, identify, track or prioritize and attack enemy targets on the battlefield.

There is growing consensus among CCW delegates and experts that the real concern, be it from a legal, ethical or security standpoint, is when autonomy is applied to the last two categories.[3] These are sometimes referred to as the 'critical functions' in weapon systems. In contrast, the functions in the first three categories are usually described as 'operational functions'. Advances in autonomy for operational functions are deemed less problematic, at least as far international law is concerned. This does not mean that these do not merit attention. Advances in autonomy in the areas of

[1] Tan, T. and Zheng, Z-Y., 'Research advances in swarm robotics', *Defence Technology*, vol. 9, no. 1 (Mar. 2013), pp. 18–39.

[2] Arquilla, J. and Ronfeldt, R., *Swarming and the Future of Conflict* (RAND Corporation: Santa Monica, CA, 2005); and Scharre, P., *Robotics on the Battlefield Part II: The Coming Swarm* (Centre for a New American Security: Washington, DC, Oct. 2014).

[3] International Committee of the Red Cross (ICRC), *Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects*, Expert Meeting Report (ICRC: Geneva, 2014); and United Nations Institute for Disarmament Research (UNIDIR), *Framing Discussions on the Weaponization of Increasingly Autonomous Technologies*, UNIDIR Resources no. 1 (UNIDIR: Geneva, 2014).

Table 1. Generic categorization of autonomous functions in military platforms and systems

General capability areas	Autonomous ability	Tasks
Mobility	Ability for the system to govern and direct its motion within its environment	Navigation Take-off/landing Collision avoidance Follow me Return to base
Health management	Ability for the system to manage its functioning and survival	Fault detection Self-repair Power management
Interoperability	Ability for the system to collaborate with other machines or humans	Multi-agent communication and coordination (swarming) Human-machine interaction through natural language communication
Battlefield intelligence	Ability to collect and process data of tactical and strategic relevance	Data collection Data analysis
Use of force	Ability to search for, identify, track or select and attack targets	Target detection Target identification Target tracking Target selection Fire control

Source: SIPRI data set on autonomy in weapon systems.

mobility, health management and interoperability could, in fact, improve the offensive potential of weapon systems. The technological developments that are enabling advances in autonomy in these functional areas may also serve to improve autonomy in the targeting process. This will be illustrated in more detail in section 5, which will present how autonomy is created in greater detail.

II. Autonomy as it applies to use of force

Systems that, once deployed, can independently execute some aspects of target detection, identification, tracking and attack do not belong to a distant future. In fact, some weapon systems with these functions have been used for decades. They include the following (see also table 2):

1. *Missile and rocket defence systems* such as the Goalkeeper close-in weapon systems developed by the Netherlands or the Iron Dome counter-rocket artillery and mortar systems developed by Israel. These are used for air defence to protect ships or ground installations against incoming missiles, rockets, artillery shells, unmanned systems or high-speed boats. Such systems use radars that can detect incoming projectiles and aircraft, and respond via computer-controlled 'fire control systems' to aim and fire interceptor missiles or bullets. These weapons have been used since the 1980s and have been deployed in at least 30 countries.

2. *Active vehicle protection weapons*, such as the Trophy system developed by Israel, are also air defence systems. They are used on armoured vehicles to protect against incoming missiles and rockets. They use fire-control radar to identify and attack incoming projectiles.

3. *Anti-personnel sentry weapons*, such as the Samsung SGR-A1 developed by the Republic of Korea, serve to protect specific locations (e.g. sites, borders etc) against intruders. They can be either fixed or mobile. The systems in use reportedly require human authorization to fire autonomously at a human target.

4. *Smart sensor-fused munitions* (such as the Bonus 155 projectile developed by France and Sweden), *guided missiles* (such as the Brimstone air-to-ground missile

Table 2. Examples of systems that select and engage targets without direct human involvement

Types of system	Examples (country of development)
Missile and rocket defence systems	Goalkeeper close-in weapon system (Netherlands) Iron Dome (Israel) Kashtan close-in weapon system (Russia)
Active vehicle protection weapons	AMAP-ADS (Germany) LEDS-150 (South Africa) Trophy (Israel)
Anti-personnel sentry weapons	Samsung SGR-A1 (Republic of Korea) Guardium (Israel) MDARS-E (USA)
Sensor-fused munitions	Bonus 155 mm projectile (France/Sweden) SMArt 155 mm projectile (Germany)
Guided missiles	Air launched: Brimstone air-to-ground missile (UK) Cruise missile: BrahMos (India/Russia) Anti-ship missile: RBS 15 (Sweden)
Loitering munitions	Harpy (Israel) Low-Cost Autonomous Attack System (USA) TARES (Germany)
Encapsulated torpedoes and mines	MK-60 CAPTOR (USA) PMK-2 encapsulated torpedo mine (Russia) Sea Urchin (UK)

Source: SIPRI data set on autonomy in weapon systems.

developed by the United Kingdom) and *loitering munitions* (including the Israeli-produced Harpy) are 'smart' munitions. These include in-built sensors and target recognition software that allow them to identify and select targets that correspond to pre-programmed target signatures. Loitering munitions feature greater autonomy than guided missiles and sensor-fused munitions as they have a high degree of freedom in terms of mobility. They can 'loiter' over a designated area during a predetermined period. Some may also have the ability to operate as a swarm.

5. *Encapsulated torpedoes and mines*, such as the MK-60 CAPTOR developed by the United States, work on the same principles as guided missiles but operate underwater. They use acoustic sensors and computer-fire control systems that allow them to recognize the signature of submarines.

Whether these systems can be described as 'autonomous' is a matter of perspective and depends on which approach to the concept of autonomy is used. From the point of view of the human-machine command-and-control relationship they are, or could be labelled, 'semi-autonomous' (some operate with human involvement or under direct human supervision). However, when assessing the decision-making capacity of the systems, the level of autonomy is more debatable. The actual capabilities and uses of these systems will be presented in greater detail in a separate working paper.

4. Unravelling the machinery

In order to understand the current state and future development of autonomy in weapon systems and military and civilian systems more generally, it is useful to describe and explain some of the technical foundations of autonomy: notably how it works and what the key enabling technologies are.

I. How does autonomy work?

From a basic technical standpoint, 'autonomy is about transforming data from the environment into purposeful plans and actions'.[1] Regardless of the nature of the human-machine relationship, the degree of sophistication of the system or the type of task that is executed, autonomy (in a physical system) is always enabled by the integration of the same three fundamental capabilities: sense, decide and act.[2] These capabilities will be presented in turn.

Sense

To complete a task autonomously a system needs to be able to perceive the environment in which it operates. For that, it requires sensors to collect data (the 'sense' part of perception) and a computer which uses a dedicated program—a sensing software—that can fuse and interpret the data (the 'think' part of perception).[3] The way sensing software works can vary significantly depending on the type of sensory data and the end use of the processed data. Many types of sensing software, notably computer vision software used for target detection, rely on pattern recognition: the software looks for predefined patterns in the raw data and compares them to example patterns stored in a computer memory, either on-board or off-board the system. It is worth emphasizing that computers identify patterns, such as for image or speech recognition, in a fundamentally different way from the way humans do. They use mathematical methods to find relationships in the sensory data. This means that when computers make errors, they are very different from those that a human would make. Recent studies have shown that state-of-the-art computer vision systems that can showcase human-competitive results on many pattern recognition tasks can easily be fooled. One study illustrated that changing an image originally correctly classified (e.g. a lion) in a way that is imperceptible to the human eye can cause the computer vision software to label the image as something entirely different (e.g. mislabelling a lion as a library).[4] Another study 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.[5]

[1] Mindell, D., *Our Robots, Ourselves, Robotics and the Myths of Autonomy* (Viking: New York, NY, 2015), p. 12.

[2] US Department of Defense (DOD), Office of Technical Intelligence, Office of the Assistant Secretary of Defense for Research and Engineering, *Technical Assessment: Autonomy* (DOD: Washington, DC, Feb. 2015), p. 2.

[3] Sensors may also be turned inwards to make the system capable of self-assessment, e.g. monitoring power resources or the state of physical components.

[4] Szegedy, C. et al., 'Intriguing properties of neural networks', arXiv:1312.6199 (2013), arxiv.org, <https://arxiv.org/pdf/1312.6199v4.pdf>.

[5] Nguyen, A., Yosinski, J. and Clune J., 'Deep neural networks are easily fooled: high prediction confidence for unrecognizable objects', Institute of Electrical and Electronics Engineers (IEEE), Computer Vision and Pattern Recognition 2015.

Decide

The data that has been processed by the sensing software serves then as input for the decision-making process, which is assured by the control system. The way the control system determines the course of action towards the task-specific goal can differ greatly from one system to another. Drawing upon Russell's and Norvig's classification of intelligent agents, two generic categories of control system (which themselves can be further divided into two types) can be identified: (*a*) reflex-based control systems (simple or model-based), and (*b*) optimization-based control systems (goal-based or utility based).[6] The decision–making processes presented by these categories differ radically from each other.

Reflex-based control systems

Reflex-based control systems can be divided into two subtypes: simple reflex-control systems and model-based reflex-control systems. 'Simple' reflex systems follow a strict sense-act modality. They merely consist of a set of condition-action rules (also known as 'if-then rules') that explicitly prescribe how the system should react to a given sensory input. To take the example of a landmine, these rules would be: *if* the weight exerted on the mine is between X and Y kilogrammes, *then* detonate. These systems succeed only in environments that are fully observable through sensors.

Model-based reflex-control systems are slightly more complex in their design as they include a 'model of the world' meaning a knowledge base that represents, in mathematical terms, how the world works: how it evolves independently of the system and how the system's actions affect it. The additional information provided by the model helps improve performance and reliability as it aids the control system to keep track of its percept history and parts of the environment it cannot observe though its sensors.[7] For instance, for an autonomous vacuum cleaner this information could simply be a map of the surface that has to be vacuumed. Like simple reflex control systems, model-based control systems follow a fixed set of rules and their decision making is implemented in some form of direct mapping from situation to action.

Optimization-based control systems

Optimization-based control systems, on the other hand, can govern their own actions by manipulating data structures representing what Weiss calls their 'beliefs', 'desires' and 'intentions'.[8] They combine (*a*) a model of the world (belief about how the world works and the reactions to the system's actions), (*b*) a value function that provides information about the desired goal (desire), and (*c*) a set of potential rules that help the system to search and plan how to achieve the goal (intention).[9] To make a decision, optimization-based control systems weigh the consequences of possible actions and measure whether and to what extent they will serve the achievement of the goal. One concrete example would be the homing function in a beyond-visual-range air-to-air

[6] Russell and Norvig define 'agents' as 'anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators'; an agent can be a human, a robot or software. A fifth subtype, a 'learning agent', could also be listed here but is discussed separately in the subsection on machine learning. Russell, S. and Norvig, P., *Artificial Intelligence: A Modern Approach*, 3rd edn (Pearson Education: Harlow, 2014), p. 35, p. 49.

[7] Russell and Norvig describe reflex agents that include a model of the world as model-based reflex agents. Those that do not have a model are referred to as a 'simple reflex agent'. Russell and Norvig (note 6).

[8] Weiss, G., *Multiagent Systems*, 2nd edn (MIT Press: Cambridge, MA, 2013), pp. 54–55.

[9] Control systems that only include goal information in their value function are counted as 'goal-based systems' under Russell's and Norvig's classification. Control systems that include information about utility of the action outcomes in their value function are called 'utility-based agents'. These agents can vector performance and efficiency factors to maximize their course of action. Utility-based agents are more intelligent and efficient than goal-based agents. They are preferable when meeting the goal cannot be achieved in a single action and the agent is required to plan a series of sequential actions. Russell and Norvig (note 6).

missile (e.g. the Meteor missile developed by the European producer MBDA). The desired goal of the missile is to attack a predetermined target. Combining input from sensors, information from the model of the world and the rules included in its utility function, the missile's control system can find the quickest and most energy-efficient route to approach the target. It can then track the target until it has an opportunity to attack it.

Optimization-based control systems feature a level of deliberative intelligence or self-governance that reflex agents do not have. They do not simply go through a series of pre-mapped actions; they can reason about the possible consequences of actions and then act accordingly. Their main advantage is flexibility. They can handle scenarios that could not be foreseen in the design stage. This does not necessarily mean, however, that their behaviour is not predictable or that the systems are capable of free will. Control systems do only what they are programmed to do, regardless of the complexity of their programming.[10]

It should be mentioned that 'randomized' algorithms can be used in both reflex-based control systems and optimization-based control systems. Randomized algorithms are 'non-deterministic' in that they allow systems to randomly pick a solution to solve a problem. In the context of a reflex-based agent, the use of randomized algorithms allows the agent to escape from an infinite loop (i.e. the situation when an agent endlessly repeats an action to meet a goal but the goal cannot be achieved by that action) by randomly picking between two predetermined alternatives. In the case of a vacuum cleaner this could be randomly turning left or right when confronted by an obstacle. In optimization-based control, the use of randomized algorithms is useful to prevent a system from having to search all possible combinations of actions. For some processes the use of random algorithms provides the simplest or fastest way to achieve a result. The use of randomized algorithms provides such systems with the potential to generate different behaviour under the same input condition. In other words, it introduces some unpredictability in the behaviour of the system. That is why the use of non-deterministic algorithms is rare in safety-critical systems.

Act

The decisions made by the control systems are then exerted in the real world through computational or physical means.[11] In the cyber-realm, for instance, this could be a software program that would implement a specific action such as blocking a malicious code. When discussing robotic platforms, the means through which the systems interact with the environment are commonly referred to as 'end-effectors' and 'actuators'. End-effectors are the physical devices that assert physical force on the environment: wheels, legs and wings for locomotion, as well as grippers and, of course, weapons. Actuators are the 'muscles' that enable the end-effectors to exert force, and include things such as electric motors and hydraulic or pneumatic cylinders. It should be noted that actuators and end-effectors might in some cases be coupled with sensors that will provide feedback information to the control systems concerning the task execution.

In summary, autonomy derives, from a technical standpoint, from the ability of a system to sense and act upon an environment and direct its activity towards achieving a given goal. Figure 1 represents in a simple fashion how these different capabilities interact with each other within a system that uses an optimization-based control system.

[10] Righetti, L., 'Emerging technology and future autonomous systems: speaker's summary', Autonomous Weapon Systems: Implication of Increasing Autonomy in the Critical Functions of Weapons, Expert Meeting, Versoix, Switzerland, 15–16 Mar. 2016, p. 39.

[11] Russell and Norvig (note 6), pp. 988–90.

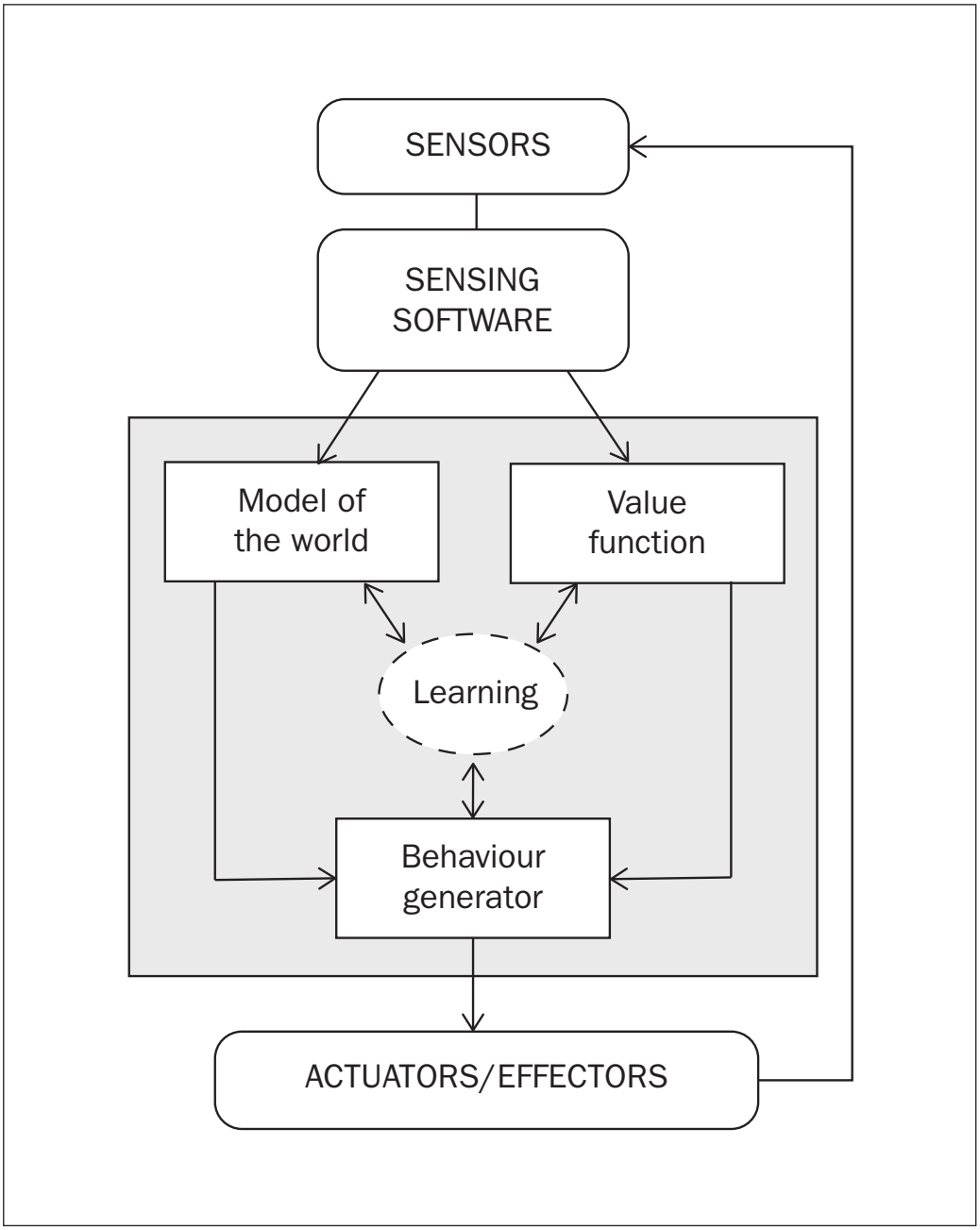

Figure 1. Anatomy of autonomy

II. What are the underlying technologies?

Anatomy of autonomy: underlying technology architecture

As implied by the previous description, autonomy is at a fundamental level always enabled by some type of underlying technology:

1. Sensors that allow the system to gather data about the world;

2. A suite of computer hardware and software that allows the system to interpret data from the sensor and transform it into plans and actions. The three most important technologies in this regard are computer chips, sensing software and control software that together form the 'brain' of the system;

3. Communication technology and human-machine interfaces that allow the system to interact with other agents, whether they be machines or humans; and

4. Actuators and end-effectors that allow the system to execute the actions in its operating environment.

These different components form the underlying architecture of autonomy. The actual characteristics of these underlying technologies will be different depending on the nature of the task and the operating environment. It should also be noted that technologies may be integrated within a single machine—which could be described as 'self-contained autonomy'—or distributed across a network of machines—which could be described as 'distributed autonomy'.

Autonomy: a 'software endeavour'

Advances in autonomy in weapon systems are dependent upon technological progress in multiple areas. Advances in sensor technologies are certainly crucial as such technologies determine the accuracy of the data that systems can collect on their operating environments. Likewise, advances in computer-processing technologies play an important role as they determine the speed at which the software part of a system can 'think' as well as the volume of data that it can efficiently handle. The design of the actuators and end-effectors will also affect the hardiness, endurance and cost of the systems.

The technologies that are deemed the most critical to autonomy, however, are the software elements. As a 2012 report by the Defense Science Board of the US Department of Defense pointed out, autonomy is primarily a 'software endeavour'.[12] It is the complexity of sensing software and control software that actually determines the level of autonomy of a system. In other words, autonomy is a very 'diffuse' technology that does not easily lend itself to being tracked or measured because it fundamentally depends on the ingenuity of human programmers to find a way to break down a problem into mathematical rules and instructions that the computer will be able to handle. That being said, the state of the art is relatively well known. The following section describes what is currently feasible for humans to achieve in programming within the bounds of contemporary knowledge.

[12] Department of Defense (DOD), Defense Science Board, *Task Force Report: Role of Autonomy in DoD Systems* (DOD: Washington, DC, 2012), p. 22.

5. Creating autonomy

This section takes stock of the extent to which autonomy remains an engineering challenge. It starts by discussing the variables that make autonomy difficult from a programming perspective. Next, it presents what is feasible with today's technology and, lastly, it discusses the extent to which the highly published progress made in machine learning could fuel significant advances in autonomy in weapon systems.

I. How difficult is it to achieve autonomy?

Achieving autonomy is, by definition, not actually that difficult. The extent to which it is feasible with today's technology depends on two interrelated variables: (*a*) the complexity of the task, and (*b*) the complexity of the environment.

The complexity of the task

The complexity of a task primarily has to do with the extent to which it is possible to model it mathematically and does not reflect how difficult its execution might be according to human standards. A famous paradox in the artificial intelligence and robotics community—known as 'Moravec's paradox'—is that 'hard problems are easy and easy problems are hard'. According to Moravec, 'it is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility'.[1]

There are several variables that contribute to making a task complex from the programmers' point of view. The first variable is precision: how well defined is the task? Does the task follow rules or a concrete logic? The more abstract or ill-defined the task specifications, the harder it is to formulate in terms of a mathematical problem and a solution. The second factor is that of tangibility: can the expected outcome be quantified? Task executions that require qualitative judgement are often problematic because the outcome cannot be assessed in objective terms. It is debatable for instance whether the principles that govern the use of force in international humanitarian law—notably proportionality and precaution in attacks—could, or should, ever be represented in terms that a computer could reason with. A third variable is dimensionality: can the task be executed in a single action or does it require sequential decisions and actions? How many possibilities are the systems facing to execute each action? The combined answers to these two questions determine the number of possibilities that the systems might have to process to take a decision. The more possibilities that exist, the more advanced the programming needs to be and the more computing power is necessary to engineer optimal solutions to a problem. A fourth variable is interaction: does execution of the task require interaction with other autonomous agents? What is the nature of the interaction: are agents competing, collaborating or simply communicating? Modelling interaction with other agents, particularly humans, in either a competitive or collaborative context is fundamentally difficult as human behaviour is often unpredictable.

[1] Pinker. S., *The Language Instinct* (Harper Perennial: New York, NY, 2007), pp. 190–91; and Moravec, H., *Mind Children* (Harvard University Press: Cambridge, MA, 1988).

The complexity of the environment

The complexity of the environment derives from several elements. Is the environment fully observable or partially observable through sensors? Is it a known or well-understood environment? Is it structured or unstructured? Is it cluttered or uncluttered? Is it static or dynamic? Is it a determinist or stochastic environment (i.e. does the system's action always produce the same effects on it?) Is it an adversarial environment where actors may actively seek to defeat the system? All these variables affect the extent to which the environment is predictable and can be modelled in advance either explicitly (like a map showing what the environment looks like precisely) or implicitly (rules about how it works, e.g. rules of the road). The less predictable the environment, the harder it is to model and therefore the harder it is to create autonomous capabilities within systems, at least those that are effective and reliable.

The case of navigational autonomy in robotic platforms provides a good illustration of the challenges posed by varying levels of complexity in different environments. Navigational autonomy is comparatively easy to create for systems operating in the air or underwater for the simple reason that generally these two domains are uncluttered: they feature a limited number of possible obstacles. In addition, the laws of physics in these two domains are well understood. Hence, they can be easily represented in mathematical terms. The land domain, on the other hand, offers greater complexity in many regards: the structure of the terrain may vary greatly, the systems may face many different types of obstacles and have to interact with other autonomous agents—either other machines or humans—whose behaviour might be unpredictable. Engineers know very well how to make self-driving vehicles that can operate within constrained and structured environments (within a factory or on the tarmac of an airport) or unpopulated or sparsely populated semi-structured environments (such as a motorway) because these can be easily explicitly mapped or implicitly modelled in advance. Making self-driving vehicles capable of operating in highly diverse human environments, such as a city centre, is much more challenging because it is difficult—if not impossible—for a programmer to develop a model that will capture all possible combinations of events. By definition, a model is a simplified version of the world; there is always a risk of a corner case (i.e. a problem or situation that has not been represented and planned for in the model of the world).

II. What is feasible with today's technology?

Presenting the current state of autonomy in a single description is difficult because the description depends upon the types of tasks and environments that are of interest. This subsection therefore focuses on what is technically possible in the five application areas of autonomy identified in table 1 (section 3): mobility, health management, interoperability, battlefield intelligence, and use of force.

Mobility

As already explained in the previous subsection, the extent to which developing autonomous navigation capabilities is feasible using current technology depends upon the complexity of the environment.

Air domain

The air domain is the environment where the interest for autonomy has been the strongest and consequently where technological progress has been the greatest. Engineers have created unmanned air systems—also known as unmanned aerial

vehicles (UAVs)—that can take off and land autonomously; fly to predetermined destinations using the Global Positioning System (GPS) and waypoint navigation; complete emergency landings; and return to base upon loss of communication. The only contexts where the development of autonomous flight capabilities remain potentially challenging (but not impossible) are in low-lying and unknown environments (due to the presence of obstacles) and contested airspace (where the use of a GPS guidance system is insufficient or impossible due to jamming). Thus, the development of vision-based guidance for UAVs has been the focus of important research in recent years. Vision-based guidance systems are image-processing systems that allow the systems to build representations of their surroundings and thereby identify obstacles. However, at present this technology is usually limited to a 'visualization' of geometry meaning that UAVs can only detect (and avoid) an obstacle's geometry. UAVs, for example, do not have the functional semantic understanding to distinguish a door from a wall; they would view both objects simply as flat planes.

Maritime domain

The maritime domain also presents relatively few difficulties for the development of autonomous navigation capabilities since it rarely features obstacles. Current unmanned surface systems are mainly operated by remote control, although some models are capable of navigating and executing preprogrammed manoeuvres autonomously using GPS and waypoint navigation sensor-based obstacle avoidance systems. Unmanned underwater systems that can navigate autonomously have existed for decades. Some of them can even operate for very long periods (several weeks or months at a time).

Land domain

For the reasons given earlier, the development of autonomous navigation capabilities in the land domain remains a major scientific challenge as far as the development of military systems is concerned.

Engineers have not yet designed robots or vehicles with sufficient perception and decision-making capabilities to respond effectively on the battlefield where environmental conditions might be unknown, unstructured, dynamic and potentially adversarial.[2] That is why current mobile ground robots developed for the military and security market are nearly all remotely operated.[3] Military unmanned ground vehicles capable of navigational autonomy, such as Israel's Guardium system, can only be used for operations such as border surveillance and control where the area is known in advance and not subject to major changes.

Health management

This subsection focuses on three specific applications of health management that have decreasing levels of technological maturity: power management, fault detection and self-repair.

Power management

Progress on the development of processes to render systems capable of managing their power resources is now fairly advanced. It is possible to develop ground, land and air systems that are capable of managing the refuelling or recharging procedure

[2] US Department of Defense (DOD), Defense Science Board, *Report of the Defense Science Board Summer Study on Autonomy* (DOD: Washington, DC, 2016), p. 13.

[3] Pomerleau, M., 'Work shifts down on driverless military convoy', Defense Systems, 31 Mar. 2016, <https://defensesystems.com/articles/2016/03/31/bob-work-autonomous-ground-vehicles-unlikely-soon.aspx>.

entirely—so long as the operational environment permits it. Cutting-edge capability in this area is exemplified by the X-47 B, a combat UAV prototype capable of autonomous aerial refuelling.[4]

Fault detection

Internal fault detection and identification is reportedly not a focus area in the development of robotics and unmanned vehicles but the technology exists.[5] The state of the art for fault detection is exemplified by the NASA (National Aeronautics and Space Administration) Deep Space One probe which used model-based detection and recovery to detect errors in software execution as well as malfunctions or damage to hardware.[6]

Self-repair

Self-repair requires both the ability to self-modify and the availability of new parts or resources to fix broken parts. Most existing physical systems lack these properties. Modular robotics is one area of robotics research that is experimenting with such capability. Modular robots consist of identical robotic modules that can autonomously and dynamically change their aggregate geometric structure to suit different locomotion, manipulation and sensing tasks. They can self-repair by detecting the failure of a module, ejecting the bad module and replacing it with one of the extra modules.[7]

Interoperability

Machine-machine teaming/swarming

Machine-machine teaming/swarming has not yet reached the point where it could be turned into a marketable application, be it in the civilian or the military sector (although it is foreseen for the coming years by experts).[8] However, the fundamentals exist; engineers know how to make large groups of robots execute simple tasks (inspection of infrastructures, surveillance of borders etc) collaboratively in the air, under the sea or on the ground. Similar to autonomous navigation, the main technical difficulties derive from the nature of the operating environment. One practical challenge in that regard is the requirement for a reliable communication infrastructure.[9] Swarm operations require a reliable communications backbone, which can be difficult to maintain (and expensive to run) in remote or communication-denied areas.

[4] 'Fueled in flight: X-47 B first to complete autonomous aerial refuelling', Navair News, 22 Apr. 2015, <http://www.navair.navy.mil/index.cfm?fuseaction=home.NAVAIRNewsStory&id=5880>.

[5] Department of Defense (DOD), Defense Science Board, *Task Force Report: Role of Autonomy in DoD Systems* (DOD: Washington, DC, 2012).

[6] Bernard, D. et al., 'Autonomy and software technology on NASA's Deep Space One', *IEEE Intelligent Systems*, vol. 10 (May/June 1999).

[7] Fitch, R., Rus, D. and Vona, M., 'A basis for self-repair robots, using reconfiguring crystal module', Institute of Electrical and Electronics Engineers (IEEE)/Robotics Society of Japan (RSJ) International Conference on Intelligent Robots and Systems 2000 (IROS 2000), Takamatsu, Japan, 30 Oct.–5 Nov. 2000, <https://groups.csail.mit.edu/drl/wiki/images/f/f9/Fitch_Rus_Vona_2000_A_Basis_for_Self-Repair_Robots_Using_Self-Reconfiguring_Crystal_Modules.pdf>.

[8] Near-term applications of swarms in the military domain include deployment of micro-drones for intelligence, surveillance and reconnaissance (ISR) missions in cluttered environments, and swarms of unmanned systems for vehicle protection and anti-access and area denial. Arquilla, J. and Ronfeldt, R., *Swarming and the Future of Conflict* (RAND Corporation: Santa Monica, CA, 2005); Scharre, P., *Robotics on the Battlefield Part II: The Coming Swarm* (Centre for a New American Security: Washington, DC, Oct. 2014); Golson, J., 'The Navy's developing little autonomous boats to defend its ships', *Wired*, 10 June 2014; and 'US military's new swarm of mini drones', *Defense News*, 17 May 2015.

[9] Tan, T. and Zheng, Z-Y., 'Research advances in swarm robotics', *Defence Technology*, vol. 9, no. 1 (Mar. 2013), pp. 18–39.

Human-machine teaming

The ultimate model of human-machine teaming for many military planners would be a situation where operators could describe and give directions—before and during operations—using natural language, and where robots or autonomous systems could report on their actions or ask for additional input or assistance when they met an unexpected situation.[10] This model is not yet achievable with current technology. Speech-interface technology has developed enormously in recent years (notably thanks to the standardization of voice-commanded digital assistants in smart phones), but it still falls short of what would be expected for a peer-to-peer human-machine communication.[11] State-of-the-art speech interfaces are steadily improving at speech recognition (recognizing words being said) but they still have major difficulties with understanding speech (recognizing what is being discussed).[12] For now, they can only handle simple queries and the fault rate remains fairly high.[13] The technology is yet to reach the point where systems can (*a*) comprehend complex spoken phrases, and (*b*) maintain an understanding of what is being discussed at an abstract level. These are two fundamental requirements for effective communication with humans.

Battlefield intelligence

Existing military systems are not yet capable of collecting and processing intelligence information independently. Surveillance UAVs have no ability to analyse information on-board; all the data that is captured must be monitored and analysed by human analysts off-board.[14] There is great interest within the military for developing systems capable of pre-process intelligence information—that is, the identification of situations of interest such as suspect human behaviour and the communication of that information to human analysts for disambiguation.[15] Arguably, state-of-the-art computer vision algorithms are sufficiently sophisticated to achieve this.[16]

Another development in autonomy worth noting (although it does not take place on-board weapon systems) is the use of big data analytics for pattern recognition in intelligence data. One recent illustration of this capability is the alleged use of machine-learning algorithms by the USA to search the Global System for Mobile (GSM) communication metadata of 55 million mobile phone users in Pakistan. The algorithm was trained to track down couriers carrying messages between Al-Qaeda members.[17]

[10] The development of human-machine communication is also seen as a way to increase human trust in autonomous systems and therefore facilitate their adoption by military personnel. US Department of Defense (DOD), Defense Science Board (note 2), p. 15.

[11] Voice recognition for command-and-control can be found in the most recent generations of combat aircraft: the F-16 Vista and the F-35 Lightning (Lockheed Martin), the JAS 39 Gripen (Saab), the Mirage (Dassault) and the Eurofighter Typhoon (Airbus). However, it is only used to operate non-critical functions. Schutte, J., 'Researchers fine-tune F-35 pilot-aircraft speech system', Air Force Link, 15 Oct. 2007, <https://web.archive.org/web/20071020030310/http://www.af.mil/news/story.asp?id=123071861>; and Englund, C., 'Speech recognition in the JAS 39 Gripen aircraft: adaptation to speech at different G-loads', Master's Thesis in Speech Technology, Royal Institute of Technology, Stockholm, 11 Mar. 2004, <http://www.speech.kth.se/prod/publications/files/1664.pdf>.

[12] Knight, W., '10 breakthrough technologies 2016: conversational interfaces', *MIT Technology Review* (2016); and Tuttle, T., 'The future of voice: what's next after Siri, Alexa and Ok Google', Recode, 27 Oct. 2015, <http://www.recode.net/2015/10/27/11620032/the-future-of-voice-whats-next-after-siri-alexa-and-ok-google>.

[13] Guo, J., 'Google new artificial intelligence can't understand these sentences. Can you?', *Washington Post*, 18 May 2016, <https://www.washingtonpost.com/news/wonk/wp/2016/05/18/googles-new-artificial-intelligence-cant-understand-these-sentences-can-you/>.

[14] Tucker, P., 'Robots won't be taking these military jobs anytime soon', Defense One, 22 June 2015, <http://www.defenseone.com/technology/2015/06/robots-wont-be-taking-these-military-jobs-anytime-soon/116017/>.

[15] The rationale behind military interest is threefold: reducing manpower burden, reducing analysis time and reducing the size of the communication broadband.

[16] US Department of Defense (DOD), Defense Science Board (note 5); and Scheidt, D., 'Organic persistent intelligence, surveillance and reconnaissance', *John Hopkins APL Technical Digest*, vol. 31, no. 2 (2012).

[17] Robbin, M., 'Has a rampaging AI algorithm really killed thousands in Pakistan?', *The Guardian*, 18 Feb. 2016.

Use of force

As previously discussed there are a number of existing weapon systems that, once deployed, can detect, select, track and engage targets without human intervention. The capabilities of automated target recognition software in terms of perception and decision making are rather limited. They can only recognize large and well-defined objects (like tanks or enemy radars), and require favourable weather conditions and uncluttered environments.[18] Their decision-making process is also highly constrained and they can only fire upon target types that have been predetermined.[19]

The capabilities of existing weapon systems are not, however, representative of what is possible in the field of pattern recognition, in particular, for image identification and classification. The field of computer vision and pattern recognition has made substantial strides in recent years notably due to the improvement of machine-learning techniques (see below).[20] It is possible to develop systems that show human-competitive results in terms of face and object recognition. In some cases such systems may even outperform humans (e.g. for facial recognition). However, it should be noted that these systems currently perform better in the cloud than in the physical world. Robotic systems need to be capable of continuous perception, which is a level of performance significantly beyond pattern recognition in images or video from the Internet. The robotics industry has only just begun to exploit the potential of the latest advances in machine learning to improve the perception capabilities of robots. Moreover, machine learning requires access to huge volumes of data, which poses the problem of data availability.

The development of neural network vision technology, which uses processes similar to those of the human brain, may provide further advances in this field. However, it is debatable whether this technology is capable of powering a next-generation target recognition system at this stage, although a number of ongoing research and development projects are pursuing that objective.[21]

Although computer vision technology has made great progress in biometrics and object recognition, it still struggles to infer abstract meaning from images, videofootage or real-life situations.[22] Cutting-edge computer vision systems can recognize some simple human actions such as walking, running and hand waving, but they are unable to determine the intentions behind these actions (e.g. why a person might be running). Making computers capable of understanding complex actions and goal-oriented activity remains a fundamental research problem. In other words, it remains challenging using the currently available technology to develop autonomous target recognition systems able to detect human enemy targets based on the behaviour or actions of those targets.

[18] Ratches, J., 'Review of current target recognition systems', *Optical Engineering*, vol. 50, no. 5 (2011), pp. 1–7; and Roff, H., 'Sensor-fused munitions, missiles and loitering munitions: speaker's summary', Autonomous Weapon Systems: Implication of Increasing Autonomy in the Critical Functions of Weapons, Expert Meeting, Versoix, Switzerland, 15–16 Mar. 2016, pp. 33–34.

[19] The performance and capabilities of existing weapons that have some level of autonomy in their critical functions will be further analysed in a separate working paper.

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

[21] A small number of large defence companies, including BAE Systems, Leidos and Lockheed Martin, are reportedly conducting research and development efforts in this area. Warwick, G. and DiMascio, J., 'Machine learning key to automatic target recognition', *Aviation Week & Space Technology*, 26 May 2016.

[22] Karpathy, A., 'The state of computer vision and AI: we are really, really far away', Andrej Karpathy Blog, 22 Oct. 2012, <http://karpathy.github.io/2012/10/22/state-of-computer-vision/>.

Conclusion

A conclusion that can be drawn from this brief review is the importance of perception. It is the lack of perceptual intelligence that is impeding the advance of autonomy in some of the most critical applications areas of autonomy in weapon systems, namely mobility, interoperability, use of force and battlefield intelligence. For a number of experts, the solution to designing machines capable of advanced situational understanding lies in the current progress of machine learning. This is the topic of the final subsection.

III. How important is machine learning to future advances of autonomy in weapon systems?

Handcraft programming vs machine learning

Currently, most software is handcrafted, meaning that human programmers are entirely responsible for defining the problems to be solved by the software and the way in which it solves those problems. This requires a great deal of research on how the world works. Engineers developing autonomous systems 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), in order to develop the model and rules that will govern the behaviour of the systems, whether for perception or decision making.

Handcraft programming has limitations, particularly when tasks and operating environments are too complex for a human to model them completely.[23] This is one of the reasons why in many areas of artificial intelligence and robotics research—two disciplines that are directly involved in the development of autonomy—programmers now rely extensively on machine learning to develop their systems.[24]

Machine learning is an approach to software development which consists of building a system that can learn and then teaching it what to do using a variety of methods (see box 1). This is a complex and data-heavy undertaking. Machines learn by abstracting statistical relationships in data. To be taught, they need to be provided with large amounts of training data (real-world examples) and rules about the data relationship. The main advantage of machine learning compared with traditional programming is that humans do not have to explicitly define the problem or the solution, instead the machine is designed to improve its knowledge through experience.

Machine learning: opportunities and challenges

Machine learning has been around for decades but has made great strides in recent years, notably due to improvements in computer power and developments in deep learning—a specific technique based on neural networks, which draws heavily on knowledge of the human brain, statistics, and applied maths (see box 2).[25] These recent advances have created both important opportunities and challenges for the development of autonomy in weapon systems.

As previously discussed, recent advances in machine learning have proved to be very useful for machine perception. They allow the programmer to design sensing software

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

[24] Russell, S. and Norvig, P., *Artificial Intelligence: A Modern Approach*, 3rd edn (Pearson Education: Harlow, 2014), p. 56.

[25] Goodfellow, I., Bengio, Y. and Courville, A., *Deep Learning*, (MIT Press: Cambridge, MA, Forthcoming); and Murnane, K., 'What is deep learning and how is it useful?', *Forbes*, 1 Apr. 2016.

Box 1. Machine-learning methods

According to Nilsson, 'a machine learns whenever it changes its structure, program, or data (based on its inputs or in response to external information) in such a manner that its expected future performance improves. Some of these changes, such as the addition of a record to a database, fall comfortably within the province of other disciplines and are not necessarily better understood for being called learning. But, for example, when the performance of a speech-recognition machine improves after hearing several samples of a person's speech, we feel quite justified in that case to say that the machine has learned'.

A machine can learn on the job (online learning) or during a training phase (offline) with a wide spectrum of methods that can be sorted into three generic categories: reinforcement learning, supervised learning and unsupervised learning.

- Reinforcement learning: the machine receives some reward for its action. It obtains more rewards when the outcome is closer to the desired outcome. This motivates it to find the most suitable solution. The desired outcome is never presented to the machine.
- Supervised learning: the machine learns by comparing example inputs with desired outputs. The data is labelled with the correct answer. Examples include systems that learn image recognition by scanning databases with tagged images.
- Unsupervised learning: the machine is only presented with raw data and it must find patterns in the data itself. It is the most difficult method of learning and the one that currently shows the least mature results.
- Semi-supervised learning: the machine is presented with both labelled and unlabelled examples of data.

In practice, the distinctions between the categories are not always clear-cut and different methods may be used to train a system.

Source: Nilsson, N. J., *Introduction to Machine Learning: An Early Draft of a Proposed Textbook* (Stanford University: Stanford, CA, 1998), p. 1.

that features remarkable capabilities in terms of pattern recognition (whether objects, faces or radio signals).[26] They create improvement opportunities in all application areas of autonomy in weapon systems, from target recognition to navigation.

Machine learning also poses a number of practical challenges, particularly with regard to predictability. Machine learning systems, particularly those that run on deep neural networks, could be said to operate like a 'black box' system: the input and output of the system are observable but the process leading from input to output is unknown or difficult to understand. It is particularly difficult for humans to understand what such systems have learned and hence how they might react to input data that is very different from that used during the training phase.[27] Likewise, unless the system's learning algorithm is frozen at the end of the training phase, once deployed, it might learn something it was not intended to learn or do something that humans do not want it to do.[28]

These are some of the reasons why the use of machine learning in the context of weapon systems has been limited to experimental research. The introduction of machine-learning capabilities in deployed systems is unlikely in the near future unless the engineer community manages to solve some of the methodological problems that learning systems, particularly those that can learn online, pose to existing methods of verification (i.e. methods that are used to ensure that a system conforms with a regulation, requirement, specification or imposed condition).

[26] Gershgorn (note 20).

[27] Postma, E., 'Deep learning: the third neural network wave', Data Science Center Tilburg Blog, Feb. 2016, <https://www.tilburguniversity.edu/research/institutes-and-research-groups/data-science-center/blogs/data-science-blog-eric-postma/>.

[28] Roff, H. and Singer P. W., 'The next president will decide the fate of killer robots—and the future of war', *Wired*, 6 Sep. 2016.

Box 2. Deep learning

Deep learning is a type of representation learning, which in turn is a type of machine learning. Machine learning is used for many but not all approaches to artificial intelligence.

Representation learning is an approach to machine learning whereby the system 'learns' how to learn: the system transforms raw data input to representations (features) that can be effectively exploited in machine-learning tasks. This obviates manual feature engineering (whereby features are hard-coded into the system by humans), which would otherwise be necessary.

Deep learning solves a fundamental problem in representation learning by introducing representations that are expressed in terms of other, simpler representations. Deep learning 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.

Deep learning was invented decades ago but has made important progress in recent years, thanks to improvements in computing power and increased data availability and techniques to train neural networks.

Source: Goodfellow, I., Bengio, Y. and Courville, A., *Deep Learning* (MIT Press: Cambridge, MA, Forthcoming), p. 8.

6. Conclusions: key takeaways for the Convention on Certain Conventional Weapons discussions

This working paper aims to clarify some of the basic aspects of autonomy and thereby provide insights for future discussions on LAWS within the framework of the CCW. The key takeaways can be summarized in three points.

Takeaway 1. Discuss advances of 'autonomy in weapon systems' rather than autonomous weapon systems or LAWS as a general category

The study of autonomy as a general attribute of a weapon system is imprecise and potentially misleading. Autonomy may serve very different capabilities in different weapon systems. For each of these capabilities the parameters of autonomy, whether in terms of the human-machine command-and-control relationship or the sophistication of the decision-making process, may vary greatly, including over the duration of a mission. In this regard, the continued reference to the concept of LAWS in the framework of the CCW is problematic. It traps states and experts into a complex and contentious discussion about the level at which a system might be deemed autonomous, while in reality the concerns—be they from a legal, ethical or operational standpoint—need to be articulated on the use of autonomy for specific functions or tasks. Future CCW discussions could, therefore, usefully benefit from a conceptual reframing and a shift from a platform- or system-centric approach to a functional approach to autonomy. Focusing on 'autonomy in weapon systems' rather than LAWS could foster a much more consensual and constructive basis for discussion.[1]

Takeaway 2. Future investigations of autonomy should not be limited to autonomy as it applies to the targeting process

There is growing agreement among CCW delegates that autonomy raises issues primarily in the context of targeting processes, whether from a legal, ethical or security standpoint. However, advances in autonomy in other functional areas should remain under scrutiny for at least two reasons. First, because some 'non-critical' applications of autonomy may actually be determinant of the offensive capability of weapon systems (think here of capabilities like navigation, swarming and self-repair), and may pose relevant concerns in terms of human control: what are the parameters of human control when robotic weapons from a large swarm? The second reason is that the technological developments that fuel advances in some functional areas, such as navigation, may also serve to improve autonomous targeting. The image processing software used to power vision-guided navigation and target recognition may share many similarities in their design.

[1] This view is also shared by a number of experts that have studied the development of autonomy in weapon systems, including Kerstin Vignard, Chief of Operations at the United Nations Institute for Disarmament Research (UNIDIR). Vignard stressed this point in her statement at the 2016 CCW Informal Meeting of Experts on Lethal Autonomous Weapon Systems in Geneva in Apr. 2016. Vignard, K., 'Statement by the United Nations Institute for Disarmament Research', 2016 CCW Informal Meeting of Experts on Lethal Autonomous Weapon Systems, Geneva, 12 Apr. 2016, <http://www.unog.ch/80256EDD006B8954/(httpAssets)/86C96CC8C7A932DCC1257F930057C0E3/$file/2016_LAWS+MX_GeneralExchange_Statements_UNIDIR.pdf>.

Takeaway 3. Consider alternative development trajectories and their potential risks

The barriers to entry toward the development of autonomous systems are very low. Most component technologies that may be used to develop autonomy are widely available in the commercial sector. The main limitation to the creation of autonomy is the ingenuity of human programmers. The risk of terrorists or criminal organizations developing low-cost autonomous weapon systems should be considered seriously.

Takeaway 4. Current advances and possible implications of machine learning deserve greater scrutiny

Recent advances in machine learning have unlocked important opportunities for advancing autonomy in weapon systems, notably in target recognition. At the same time, these advances pose new regulatory challenges. One key practical problem is the question of how the behaviour of offline and online learning systems should be regulated and controlled through validation and verification procedures.

Appendix A: Existing definitions of autonomous weapon systems

Broadly speaking the definitions of autonomous weapon systems can be classified into three groups.

The first category consists of definitions that are articulated around the nature of the human-machine command-and-control relationship. It includes the definition supported by the United States, which describes 'autonomous weapon systems' as 'a weapon that, once activated, can select and engage targets without further intervention by a human operator'.[1] It also encompasses the definition proposed by Human Rights Watch (HRW), the non-governmental organization that coordinates the International Campaign to Stop Killer Robots. HRW makes a distinction between human-in-the-loop weapons, human-on-the-loop weapons and human-out-of-the-loop weapons. Human-out-of-the-loop weapons are robots that are capable of selecting targets and delivering force without any human input or interventions.[2]

The second category includes definitions that are based on capability parameters. The United Kingdom's definition, for instance, defines 'weapon systems' as systems 'capable of understanding higher level intent and direction. From this understanding and its perception of its environment, such a system is able to take appropriate action to bring about a desired state. It is capable of deciding a course of action, from a number of alternatives, without depending on human oversight and control, although these may still be present'.[3] Canada likewise states that 'autonomy is a subjective assessment of a robot's capabilities given the demands of mission, environment, and mechanical system'.[4]

The definitions in the third categories are structured along legal lines and lay emphasis on the nature of tasks that the systems perform autonomously. The definition favoured by the International Committee of the Red Cross presents 'autonomous weapons' as an umbrella term that would encompass any type of weapon with 'autonomy in its "critical functions", meaning a weapon that can select (i.e. search for or detect, identify, track) and attack (i.e. intercept, use force against, neutralise, damage or destroy) targets without human intervention'.[5] Switzerland's working definition describes 'autonomous weapon systems' as 'weapons systems that are capable of carrying out tasks governed by IHL [international humanitarian law] in partial or full replacement of a human in the use of force, notably in the targeting cycle', although it explicitly states that this should not necessarily be limited to the targeting cycle.[6]

This classification of definitions is, of course, barely an ideal type and does not cover all definitions. The Holy See, for example, uses a mixture of definitions characterizing armed autonomous robots using '(1) the degree and duration of supervision, (2) the

[1] United States Department of Defense, Directive 3000.09 on Autonomy in Weapon Systems, 21 Nov. 2012, <http://www.dtic.mil/whs/directives/corres/pdf/300009p.pdf>, pp. 13–14.

[2] Docherty, B., *Losing Humanity: The Case Against Killer Robots* (Human Rights Watch/International Human Rights Clinic: Washington, DC, 2012).

[3] British Ministry of Defence (MOD), Development, Concepts and Doctrine Centre (DCDC), *Joint Doctrine Note 2/11: The UK Approach to Unmanned Aircraft Systems* (MOD DCDC: Shrivenham, 30 Mar. 2011).

[4] Government of Canada, 'Canadian food for thought paper: mapping autonomy', [n.d.], CCW Informal Meeting of Experts on Lethal Autonomous Weapon Systems, Geneva, 11–15 Apr. 2016, <http://www.unog.ch/80256EDD006B8954/(httpAssets)/C3EFCE5F7BA8613BC1257F8500439B9F/$file/2016_LAWS+MX_CountryPaper+Canada+FFTP1.pdf>.

[5] International Committee of the Red Cross (ICRC), 'Autonomous Weapons: is it morally acceptable for a machine to make life and death decisions?', Statement of the ICRC at the CCW Meeting of Experts on Lethal Autonomous Weapon Systems, Geneva, 13–17 Apr. 2015, <https://www.icrc.org/en/document/lethal-autonomous-weapons-systems-LAWS>.

[6] Government of Switzerland, 'Towards a "compliance-based" approach to LAWS', Informal Working Paper, 30 Mar. 2016, CCW Informal Meeting of Experts on Lethal Autonomous Weapon Systems, Geneva, 11–15 Apr. 2016, <http://www.unog.ch/80256EDD006B8954/(httpAssets)/D2D66A9C427958D6C1257F8700415473/$file/2016_LAWS+MX_CountryPaper+Switzerland.pdf>.

predictability of the behaviour of the robot, (3) and the characteristics of the environment in which it operates'.[7]

[7] Holy See, 'Element supporting the prohibition of LAWS', Working Paper, 7 Apr. 2016, CCW Informal Meeting of Experts on Lethal Autonomous Weapon Systems, Geneva, 11–15 Apr. 2016, <http://www.unog.ch/80256EDD006B8954/(httpAssets)/752E16C02C9AECE4C1257F8F0040D05A/$file/2016_LAWSMX_CountryPaper_Holy+See.pdf>.

sipri

STOCKHOLM INTERNATIONAL
PEACE RESEARCH INSTITUTE