2014 6th International Conference on Cyber Conflict
P.Brangetto, M.Maybaum, J.Stinissen (Eds.)
2014 © NATO CCD COE Publications, Tallinn

Artificial (Intelligent) Agents and Active Cyber Defence: Policy Implications

Caitríona H. Heinl
Research Fellow
Centre of Excellence for
National Security (CENS)
S. Rajaratnam School of International Studies
Singapore

Abstract: This article examines the implications of employing artificial (intelligent) agents for active cyber defence (ACD) measures, in other words proactive measures, in the context of military and private sector operations. The article finds that many complex cyber-related challenges are solved by applying artificial intelligence (AI) tools, particularly since intelligent malware and new advanced cyber capabilities are evolving at a fast rate and intelligent solutions can assist in automation where pre-fixed automation designs are insufficient. Intelligent agents potentially underpin solutions for many current and future cyber-related challenges and AI therefore plays a possible role as one of a number of significant technical tools for ACD. However, this article considers that although such advanced solutions are needed, it finds that many technical and policy-related questions still surround the possible future consequences of these solutions, in particular the employing of fully autonomous intelligent agents and possible disruptive technologies that combine AI with other disciplines. While these AI tools and ACD actions might be technologically possible, the article argues that a number of significant policy gaps arise such as legal question marks, ideological and ethical concerns, public perception issues, public-private sector ramifications, and economic matters. It highlights several areas of possible concern and concludes that it is important to examine further the implications of these rapidly evolving developments. Finally, the article provides several policy options as a start so as to begin responsibly shaping the future policy landscape in this field.

Keywords: *artificial intelligence, artificial (intelligent) agents, active cyber defence, autonomy*

1. INTRODUCTION

Given that current cyber defence measures, in particular passive cyber defences, are inadequate for increasingly sophisticated threats, many argue for proactive measures to be taken. This article therefore examines the implications of employing artificial (intelligent) agents for active cyber defence (ACD) measures in the context of military and private sector operations.

The article finds that many cyber-related challenges are solved by applying artificial intelligence (AI) tools, particularly since intelligent malware and new advanced cyber capabilities are evolving at a rapid rate. Employing AI techniques and intelligent solutions for the purposes of dealing effectively with complex cyber-related threats is then best explained by the ability of these technologies to assist in automation since pre-fixed automation designs are insufficient. Intelligent agents potentially underlie solutions for many current and future cyber-related challenges and AI therefore plays a possible position as one of a number of significant technical tools for ACD.

However, this article argues that although such advanced solutions are required, many technical questions and uncertainties still surround the possible future consequences of their use, most particularly for the employing of fully autonomous intelligent agents and possible disruptive technologies that combine AI with other disciplines. Therefore, while numerous AI applications are already in use for cyber-related issues, this article suggests that the potential policy implications of a number of emerging and proposed techniques including possible disruptive technologies now require serious consideration. Although these AI tools and ACD actions might be technologically possible, the article considers that there are a number of serious legal implications, ideological and ethical concerns, public perception issues, public-private sector ramifications, and economic matters that could arise. It finds that to date, insufficient widespread attention has been paid in the public policy domain to many of these gaps in policy. The article concludes that there is a significant time-sensitive need to commence an in-depth further examination and serious public discourse on these issues in order to develop the future policy landscape, and finally, it provides several possible policy options that could be considered.

The article is organised as follows:

- Section 2 explores the core background concepts of artificial intelligence.
- Section 3 outlines cyber-related challenges for which AI solutions could be effectively employed.
- Section 4 considers active cyber defence and the possible roles of AI.
- Section 5 examines potentially successful emerging AI technologies.
- The final section discusses several possible policy implications based on the findings of this article and provides a number of policy recommendations.

2. BACKGROUND: CORE AI CONCEPTS

AI or computational intelligence is generally defined as technology and a branch of computer science that develops intelligent machines and software. It is regarded as the study of the design of intelligent agents where an intelligent agent is a system that perceives its environment and takes actions to maximise its chances of success. Intelligent agents are software components with features of intelligent behaviour such as (at a minimum) pro-activeness, the ability to communicate, and reactivity (in other words the ability to make some decisions and to act).[1]

[1] Enn Tyugu, "Command and Control of Cyber Weapons", *4th International Conference on Cyber Conflict*, Tallinn, 2012.

Additionally, AI may be described as the automation of activities such as decision-making, problem solving, learning, and the study of the computations that make it possible to perceive, reason, and act. It can assist planning, learning, natural language processing, robotics, computer vision, speech recognition, and problem solving that requires large amounts of memory and processing time. And while AI may be considered as a science for developing methods to solve complex problems that require some intelligence such as making the right decisions based on large amounts of data, it may also be viewed as a science that aims to discover the essence of intelligence and develop generally intelligent machines.[2] General intelligence is predicted by some to come into being by 2050, possibly leading to singularity, in other words the technological creation of intelligence superior to human intelligence. Approaches for improving machine intelligence are progressing in areas such as the expression of emotion, language interaction, as well as face recognition and forecasts suggest that they will be "interim substitutes" before direct machine intelligence is realised but for now a further maturation of AI techniques and technologies is required.[3]

Several examples of AI in use include Deep Blue (IBM's chess playing computer), autonomous vehicles that drive with traffic in urban environments[4], IBM's Watson (the computer system that can answer natural language questions), and the X-47 robotic aircraft which recently landed autonomously.[5] In addition, although not readily apparent to those working outside the field, many AI technologies such as data mining or search methods are part of everyday use. This phenomenon, where a technique is not considered as AI by the time it is used by the general public, is described as the "AI effect". It is a particularly significant concept in that public perception of what constitutes AI as well as acceptance of these tools, especially the more advanced future tools, could play an important role in the shaping of future policies. Some well known examples of the AI effect include Apple's Siri application which uses a natural language user interface to answer questions and make recommendations, Google's new Hummingbird algorithm which makes meaning of the search query for more relevant "intuitive" search results, and Google's self-driving cars.

Employing AI technologies and techniques for the purposes of cybersecurity, cyber defence (or cyber offence) and ACD is currently best explained by the ability to assist in automation. Many contend that automation is essential for dealing effectively with cyber-related threats and that many cyber defence problems can only be solved by applying AI methods. Intelligent malware and new advanced cyber capabilities are evolving rapidly, and experts argue that AI can provide the requisite flexibility and learning capability to software.[6] Intelligent software is therefore being increasingly used in cyber operations and some argue that cyber defence systems could be further adaptive and evolve dynamically with changes in network conditions

[2] Enn Tyugu, "Artificial Intelligence in Cyber Defense", *3rd International Conference on Cyber Conflict*, Tallinn, 2011.

[3] Development, Concepts and Doctrine Centre (DCDC), UK Ministry of Defence, *Strategic Trends Programme: Global Strategic Trends – Out to 2040*, 4th ed., January 2010.

[4] Defense Advanced Research Projects Agency (DARPA), United States, "DARPA Urban Challenge", http://archive.darpa.mil/grandchallenge/, November 2007.

[5] Alessandro Guarino, "Autonomous cyber weapons no longer science-fiction", Engineering and Technology Magazine, Vol 8 Issue 8, http://eandt.theiet.org/magazine/2013/08/intelligent-weapons-are-coming.cfm, 12 August 2013.

[6] Tyugu, Artificial Intelligence in Cyber Defense.

by implementing dynamic behaviour, autonomy, and adaptation such as autonomic computing or multi-agent systems.[7]

3. CYBER-RELATED CHALLENGES: AI SOLUTIONS

Although many AI methods are currently available for cyber defence, there is still an identified need for further advanced solutions, intelligent decision support, automated knowledge management and rapid situation assessment[8] for the more complex cyber-related problems. In short, reports state that intelligent systems and networks, even self-repairing networks, could increase resilience in the longer term.[9] Pre-fixed automation designs are not sufficiently effective against evolving cyber incidents for instance. New vulnerabilities, exploits and outages can occur simultaneously and at any point in time,[10] and experts contend that it is difficult for humans to effectively handle the sheer volumes of data and speed of processes without high degrees of automation - very fast, if not automated, reaction to situations, comprehensive situation awareness, and a handling of large amounts of information at a rapid rate to analyse events and make decisions is therefore considered necessary.[11]

A recent United States Department of Defense report[12] explains that the identification of operationally introduced vulnerabilities in complex systems is extremely difficult technically, and "[i]n a perfect world, DoD operational systems would be able to tell a commander when and if they were compromised, whether the system is still usable in full or degraded mode, identify alternatives to aid the commander in completing the mission, and finally provide the ability to restore the system to a known, trusted state. Today's technology does not allow that level of fidelity and understanding of systems." The report then outlines the need for the development of capacity to conduct "many, potentially hundreds or more, simultaneous, synchronized offensive cyber operations while defending against a like number of cyber attacks". For now however, it describes system administrators as inadequately trained and overworked, a lack of comprehensive automation capabilities to free personnel for serious problems, and an inadequate visibility into situational awareness of systems and networks. In addition, systems such as automated intrusion detection, automated patch management, status data from each network, and regular network audits are currently unavailable.

[7] Igor Kotenko, "Agent-based modelling and simulation of network cyber-attacks and cooperative defence mechanisms", St. Petersburg Institute for Informatics and Automation, Russian Academy of Sciences, available at: http://cdn.intechopen.com/pdfs/11547/InTech-Agent_based_modeling_and_simulation_of_network_infrastructure_cyber_attacks_and_cooperative_defense_mechanisms.pdf, 2010.

[8] Tyugu, Artificial Intelligence in Cyber Defense.

[9] DCDC, *Global Strategic Trends*.

[10] Beaudoin, Japkowicz & Matwin, "Autonomic Computer Network Defence Using Risk State and Reinforcement Learning", Defense Research and Development Canada, 2012.

[11] Tyugu, Artificial Intelligence in Cyber Defense.

[12] Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, *Resilient Military Systems and the Advanced Cyber Threat*, United States Department of Defense, Defense Science Board, January 2013.

Intelligent agents and AI-enhanced tools potentially play a significant role by underpinning solutions for several, if not most, of these problems as well as the following cyber-related challenges:[13]

- The need for continual collection, comprehensive understanding, analysis and management of large amounts of dynamic data, in other words knowledge management, from a plethora of sources and devices to develop actionable intelligence.
- Insufficient pattern recognition and behavioural analysis across different data streams from many channels.
- Lack of visibility of the complete conditions of the IT environment, and insights into possible threats and systems compromise in real time.
- Non-identification of unusual behaviour, systems and network traffic, in other words anomalies, and unusual user behaviour to spot insider threats and internal misuses.
- The need for comprehensive knowledge of the threats for decision support and decision-making.
- Intrusion detection.
- Situational awareness and continual monitoring so as to detect and mitigate attacks.
- Harnessing of information to prevent, detect and even "predict" (or rather foresee) attacks.
- Insufficient passive defences and resilience of systems to attacks.

Lastly, one of the core challenges facing nations and corporations today includes the difficulties in identifying, training and retaining skilled individuals and general consensus currently holds that the numbers working in this area need to markedly increase. However, recent defence reports from the U.S. now identify that there is a "burnout factor beginning to exhibit itself"[14] among the current cyber workforce. Therefore, although increasing the number of "cyber warriors" might alleviate the current cybersecurity skills gap to a certain degree, AI and advanced automation of particular tasks could be highly beneficial over the longer term. Furthermore, strains on labour and financial resources might be alleviated. This issue therefore requires serious consideration and further concrete analysis, especially in light of future expected trends in demographics, which according to some defence officials will work against several countries.[15]

4. ACTIVE CYBER DEFENCE AND INTELLIGENT AGENTS

> *But this virtual version of vigilante justice is fraught with peril….*[16]

[13] General information from: Security for Business Innovation Council, "Getting Ahead of Advanced Threats: Achieving Intelligence-Driven Information Security", RSA Sponsored Report, 2012; and Mirko Zorz, "Complex security architectures and innovation", http://www.net-security.org/article.php?id=1692&p=1, 29 March 2012.

[14] Under Secretary of Defense, *Resilient Military Systems*.

[15] William Lynn III, former United States Under Secretary of Defense, "2010 Cyberspace Symposium: Keynote – DoD Perspective", 26 May 2010.

[16] Gregory Zeller, "Cyber warriors eye move to 'active defense'", Long Island Business News, 25 February 2013.

Current defence measures are not considered as prepared for the limitless ways to attack a network,[17] and many argue that passive defence alone may not be sufficient.[18] Arguments are therefore being made for policy makers and network defenders to incorporate lessons such as “the best defence includes an offence”, in other words active cyber defence. William Lynn III, former United States Under Secretary of Defense, argues for instance[19] that in cyber, offence is dominant and “we cannot retreat behind a Maginot Line of firewalls” - defences should therefore be dynamic and responses at network speed as attacks happen or before they arrive. Corporations and government bodies are beginning to use ACD techniques more frequently, and this section therefore explores those aspects of ACD where AI could play a role as one of a number of technical tools in the ACD toolbox.

Although there is no universal definition for the term, for the purposes of this article ACD is understood to entail proactive measures that are launched to defend against malicious cyber activities. According to a recent CNAS analysis[20] on ACD options available to the private sector, one of the few formal definitions is found within the United States 2011 Department of Defense Strategy for Operations in Cyberspace: “DoD’s synchronized real-time capability to discover, detect, analyze, and mitigate threats and vulnerabilities. It builds on traditional approaches to defending DoD networks and systems, supplementing best practices with new operating concepts. It operates at network speed by using sensors, software, and intelligence to detect and stop malicious activity before it can affect DoD networks and systems. As intrusions may not always be stopped at the network boundary, DoD will continue to operate and improve its advanced sensors to detect, discover, and mitigate malicious activity on DoD networks.”

The CNAS analysis lays out a framework (adapted in Figure 1 below) to show that it is at the Delivery phase, during the Cyber Engagement Zone, that employing ACD techniques becomes most significant, in other words when the defender can take the initiative. However, organisations are often unaware of a compromise until the Command and Control (C2) phase when installed malware communicates outside the organisation under attack. Under this analysis, three ACD concepts are identified for responding to an attack: detection and forensics, deception, and attack termination. For detection, a number of ACD techniques to detect attacks that circumvent passive defences may be used, and once information is gathered it can inform the company’s response decisions. Detection can be by way of local information gathering using ACD techniques within the organisation’s networks, or by what is known as remote information gathering where an organisation may gather information about an incident outside its own networks (by for example accessing the C2 server of another body and scanning the computer, by loading software, removing or deleting data, or stopping the computer from functioning). For attack termination, ACD techniques can stop an attack while it is occurring by, for instance, preventing information from leaving the network or by stopping the connection between the infected computer and the C2 server. More aggressive actions could include “patching computers outside the company’s network that are used to launch attacks, taking

[17] David T. Fahrenkrug, Office of the United States Secretary of Defense, “Countering the Offensive Advantage in Cyberspace: An Integrated Defensive Strategy”, *4th International Conference on Cyber Conflict*, Tallinn, 2012.

[18] Porche, Sollinger & McKay, “An Enemy Without Borders”, U.S. Naval Institute Proceedings, October 2012.

[19] Lynn, 2010 Cyberspace Symposium.

[20] Irving Lachow, “Active Cyber Defense: A Framework for Policymakers”, Center for a New American Security, February 2013.

control of remote computers to stop attacks, and launching denial of service of attacks against attacking machines.”

While ACD actions such as deploying honeypots, actively tracking adversaries’ movements, using deception techniques, watermarking documents and terminating connections from the C2 node to infected computers do not seem to be illegal, the CNAS study concludes that there is an absence of clear national and international law for some actions, particularly remote information gathering and some of the more aggressive actions. In effect, ACD options that involve retaliation or “hacking back” are generally considered illegal (whether the ACD response is before, during or after an incident) since attempts are made to access the systems of another organisation without permission so as to access or alter information on the C2 server or computers. The study further finds that it is unclear whether accessing the C2 server of another organisation could violate privacy laws and expose a company to civil actions as well as criminal prosecution. In addition, if an organisation is in another jurisdiction, a company could possibly violate that country’s national laws, even if not violating its own. It is also unclear whether a company could legally patch the C2 server of another organisation since it would entail altering or deleting information on its computers. Finally, when the C2 server is not directly connected to the adversary but “several hops away”, not only is it technically challenging to find the source of the attacks but the company tracing the sources could violate its own national laws, those of multiple other jurisdictions, and international laws such as the Budapest Convention on Cybercrime.

FIGURE 1: CYBER KILL-CHAIN (ADAPTED FROM LACHOW, “ACTIVE CYBER DEFENSE: A FRAMEWORK FOR POLICYMAKERS”, CNAS, 2013)

Phase	Description
Reconnoiter/ Reconnaissance	Adversary researches, identifies and selects its targets.
Weaponise	Adversary couples malware with a delivery mechanism, often using an automated tool.
-	Cyber Engagement Zone:
Deliver	Adversary transmits weaponised payload to the target through emails or websites for example.
Exploit	Malware delivered to the target is triggered when the user takes an action such as opening email attachments or visiting an infected site.
Install	The malware infects the user’s system. It may hide itself from malware detection software on that system.
Command and Control (C2)	The malware sends an update on its location and status to a C2 server, often through encrypted channels that are hard to detect.
Act	The malware takes actions for the adversary such as exfiltrating, altering or destroying data.

This framework is a helpful tool to clarify when AI techniques might play a significant role. For instance, the time between an attack and systems compromise can often take minutes yet it could take months to discover the breach.[21] AI techniques could therefore be of particular value in these earlier phases of the Cyber Engagement Zone. They can assist earlier detection of compromise and provide situational awareness. In particular since active defence demands high levels of situational awareness to respond to the threat of intrusion.[22] They can also assist information gathering and decision support. Deception techniques such as proposals for experimental frameworks for autonomous baiting and deception[23] of adversaries could also be useful.

However, although these ACD concepts are technologically possible, there is legal uncertainty and it is therefore unclear whether AI tools could (or should) be used as possible ACD techniques. Before employing these tools for ACD actions, legal certainty should therefore be sought so that existing laws are not violated, even where it might be argued that the law is "grey" or national and international law is unclear.

5. CYBER GAME CHANGERS: EMERGING EFFECTIVE INTELLIGENT AGENTS & AI COMBINED WITH OTHER DISCIPLINES

While numerous AI applications such as neural networks, expert systems, intelligent agents, search, learning, and constraint solving are in use for several cyber-related challenges, a number of emerging and proposed intelligent agent hybrid technologies and techniques require further research and consideration (for example, agent-based distributed intrusion detection and hybrid multi-agent/neural network based intrusion detection). Most particularly, the policy ramifications of possible future tools that combine AI technologies with other disciplines should be seriously analysed since these tools could prove to be disruptive technologies and cyber game changers if successfully developed in the medium to long term. Further research should therefore be conducted in the near term on the consequences of their possible development.

A recent analysis of the future strategic context for defence to 2040 by the Development, Concepts and Doctrine Centre (DCDC) of the UK Ministry of Defence[24] states that advances in robotics, cognitive science coupled with powerful computing, sensors, energy efficiency and nano-technology will combine to produce rapid improvements in the capabilities of combat systems. The report explains that advances in nanotechnology will underpin many breakthroughs and that developments in individual areas are likely to be evolutionary. However, developments may be revolutionary where disciplines interact, such as the combination of cognitive science and ICT, to produce advanced decision-support tools. Furthermore, according to this report, research on mapping or "reverse engineering" the human brain will likely lead to development of "neural models" and this combined with other systems such as sensors may provide human like qualities for machine intelligence. The simulation of cognitive processes using AI is likely

[21] Costin Raiu, Kaspersky Labs, "Cyber Terrorism – An Industry Outlook", Cyber Security Forum Asia, 03 December 2012.

[22] Fahrenkrug, Countering the Offensive Advantage.

[23] Bilar & Saltaformaggio, "Using a Novel Behavioural Stimuli-Response Framework to Defend against Adversarial Cyberspace Participants", *3rd International Conference on Cyber Conflict*, Tallinn, 2011.

[24] DCDC, *Global Strategic Trends*.

to be focused in the short term on probability and pattern recognition and in the longer term to aid knowledge management and support decision-making.

In light of several conclusions within the DCDC report,[25] and for the purposes of this article, the possible future consequences of the following disciplines and technologies should be seriously considered from a policy perspective:

- *Quantum Computing:* Processing capabilities could possibly increase by 100 billion times.
- *Simulation:* Advances in mathematical modelling, behavioural science and social science will seemingly combine for more informed decision-making while advances in processing techniques and computational power will allow more comprehensive modelling and potentially enable better pattern recognition.
- *Virtual Databases:* Development of the semantic web and associated technologies will create an integrated data store with unprecedented level of access that could be exploited by reasoning techniques for more sophisticated analysis that may expose previously unseen patterns with potentially unforeseeable consequences. Sophisticated data mining tools will include automatic data reduction/filtering and automated algorithmic analysis for faster access to relevant information. "Virtual Knowledge Bases" will apparently store knowledge within large database structures in formats that intelligent software could use for improved searching, to answer questions across the whole knowledge store in near natural language form, and to issue automated situation reports on demand or in response to events to assist situational awareness.
- *Cognitive and Behavioural Science:* Certain advances such as neuro-imaging technologies may make mapping of brain activity with behaviour more reliable. Modelling techniques are likely to become more powerful and capable of more accurately understanding the complexity of human behaviour and performance which could lead to an ability to "map the human terrain".

  Advancing the field of brain sciences could open opportunities for new means to develop AI and studies are being conducted to understand the brain and how human brain function could be used as a framework for improving technologies such as cybersecurity and mobile security technologies - for example, cognitive security technology modelled after human brain function for the next generation of technology security.[26] Further, a reported new trend is the application of AI and cognitive methods in situation awareness which permits fusion of human and computer situation awareness, and supports real time and automatic decision-making.[27]

  However, commentators also contend that AI is not yet, and may never be, as powerful as "intelligence amplification", in other words when human cognition is augmented by close interaction with computers.[28] For example, after Deep Blue beat Kasparov, he tested what would happen if a machine and human chess player were paired in collaboration and found that human-machine teams, even when they did not

[25] DCDC, *Global Strategic Trends*.

[26] Center for Systems Security and Information Assurance, Cyber Defense and Disaster Recovery Conference 2013: Mobile Security.

[27] Tyugu, Command and Control of Cyber Weapons.

[28] Walter Isaacson, "Brain gain?", Book Review of Smarter Than You Think by Clive Thompson, International New York Times, 2-3 November 2013.

include the best grandmasters or most powerful computers, consistently beat teams composed solely of human grandmasters or computers.[29]

- *Autonomous Systems and Robotics:* Growth in the role of unmanned, autonomous and intelligent systems is expected. These systems could range from small sensors and personalised robots replicating human behaviour and appearance to a "cooperative plethora of intelligent networks or swarms of environmental-based platforms with the power to act without human authorisation and direction"[30] with a range of autonomy from fully autonomous to significantly automated and self-coordinating while still under high-level human command.

  Although software with intelligent agent characteristics is already in use, both technical and policy-oriented research should be further conducted on the possible consequences of employing fully autonomous intelligent agents. Autonomous intelligent agents are defined as "systems situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future - the agent is strictly associated with its environment, in other words it can be useless outside the environment for which it was designed or not an agent at all".[31]

  According to Guarino,[32] they can be purely software operating in cyberspace (computational agents) or integrated into a physical system (robotic agents) where they underpin the robot's behaviour and capabilities. Computational autonomous agents could be used for intelligence-gathering and military operations, in particular during the Reconnaissance phase for automatic discovery of vulnerabilities in target systems for example or for gathering intelligence. Autonomous agents could then develop ways to exploit these vulnerabilities and they will not need fixed and pre-programmed methods to penetrate the target system since they will analyse the target, autonomously select the points of vulnerability, and develop means to use these points so as to infiltrate the system. Currently however these capabilities are manually developed or bought on the open market since full automation of exploit development is still not widely available. Guarino continues, that although an agent's goals and targets could be pre-programmed and precisely stated to facilitate its task and to ensure legality, it could in fact occur that sometimes it might be deemed preferable to give the agent "free rein".

  The Command and Control (C2) phase therefore presents significant difficulties and warrants further attention, particularly since command and control could be hard to achieve. Experts warn that the more intelligent software becomes, the more difficult it could be to control and the C2 phase causes new threats that are difficult to avoid due to the complexity of the agents' behaviour, in particular its misunderstanding a situation, misinterpretation of commands, loss of contact and formation of unwanted coalitions, unintentionally behaving in a harmful way or its unexpected actions and unpredictable behaviour.[33]

[29] Isaacson, Brain gain?

[30] DCDC, *Global Strategic Trends*.

[31] Alessandro Guarino, "Autonomous Intelligent Agents in Cyber Offence", *5th International Conference on Cyber Conflict*, Tallinn, 2013.

[32] Guarino, Autonomous Intelligent Agents.

[33] Tyugu, Command and Control of Cyber Weapons.

6. UNCERTAIN POLICY RAMIFICATIONS

To Every Man is Given the Key to the Gates of Heaven.
The Same Key Opens the Gates of Hell.[34]

These possible developments raise significant unanswered questions and concerns. At this juncture however, technical and policy-oriented solutions, at least those in the public domain, are sparse. Concrete efforts to further clarify these gaps should therefore be conducted as soon as possible, with particular focus on ideological and ethical concerns, public perception, the interplay between the public and private sectors, economic matters, and legal implications that could arise. It is pertinent that further analysis be conducted without delay so as to develop and implement, where possible, both policy-based solutions and technological safeguards from the outset.

Suffice to say that the "Internet of the Future" will not look like the Internet of today and further challenges will also include the Internet of Things and unanticipated new usages.[35] Like previous inventions, strategic reports foresee that many of these technological developments could have positive consequences, including unintended, but some could also present threats or have "catastrophic effects".[36] In particular, these reports outline[37] that reliance on AI could create new vulnerabilities that could be exploited by adversaries and there is a high chance that malicious states and non-state actors could acquire such capabilities. Further attention should therefore focus on how this threat could be thwarted and what possible technological or policy-oriented solutions could be found to mitigate malicious applications of these future tools.

Advanced intelligent systems could also challenge the interaction between automated and human components, and the complexity of controlling multiple autonomous systems and interpreting information could become extremely difficult. Forecasts suggest that those unable for these challenges may be replaced by intelligent machines or "upgraded" by technology augmentation. Autonomic defences might even be developed to take over when human judgement is deemed "too affected by emotions or information overload".[38]

A number of technical recommendations[39] so far suggested include ensuring in the design and development of new intelligent "cyber weapons" that 1) there is a guarantee of appropriate control over them under any circumstances; 2) strict constraints on their behaviour are set; 3) they are carefully tested (although thorough verification of their safety and possible behaviours is apparently difficult); and 4) the environment is restricted as much as possible by only permitting the agent to operate on known platforms. Questions such as to what extent an agent could communicate with its "base", and whether communication should be one-way (intelligence gathering from the agent for instance) or two-way in that the C2 structure could

[34] Richard P. Feynman, "The Pleasure of Finding Things Out: The Best Short Works of Richard P. Feynman, 1999.

[35] Golling & Stelte, "Requirements for a Future EWS – Cyber Defence in the Internet of the Future", *3rd International Conference on Cyber Conflict*, Tallinn, 2011.

[36] DCDC, *Global Strategic Trends*.

[37] DCDC, *Global Strategic Trends*.

[38] Bilar & Saltaformaggio, Novel Behavioural Stimuli-Response.

[39] Tyugu, Command and Control of Cyber Weapons.

issue instructions including target selection or self-destruct commands[40] should also be further examined. Particular attention should also be drawn to dealing with the possible cooperative behaviour of agents, in other words what is described as the “multi-agent” threat.

Tyugu[41] explains that since agents can be used most efficiently in multi-agent formations, it is expected that this will be the main form of agent application in cyber operations. They could for instance negotiate between themselves and cooperatively create a complex behaviour for achieving the general goals stated by a commander but this apparently means that the strict control of behaviour of each single agent will be weaker and it will be impossible to verify the outcome of multi-agent behaviour for all situations. He explains that unwanted coalitions could possibly occur if agents have too much autonomy in decision-making since communication between agents will only be partially visible to human controllers (Guarino argues that this could be extremely difficult to disable[42]). Technical solutions recommended for these problems so far include building safeguards such as backdoors and forced destruction into agents or self-destruction if loss of contact occurs.

Further clarity and certainty on these questions should however be sought as well as on the possible legal implications where recent analyses conclude that there is a certain amount of uncertainty. Under Guarino’s analysis,[43] autonomous agents are similar to any other tool or cyber weapon employed and therefore fall under existing international law but it is unclear whether a creating state could always be held responsible if an agent exceeds its assigned tasks and makes an autonomous decision. For instance, for attribution purposes, the creators might not have known in advance the precise technique employed or the precise system targeted. Guarino therefore recommends the identification of autonomous agents, perhaps through mandatory signatures or watermarks embedded in their code, and the possible revising of international law. Lastly, if a fully autonomous agent is used as a weapon in self-defence, he also recommends that care be taken in the C2 function to clearly state the agent’s targets and build in safeguards.

However, although technical safeguards such as mandatory signatures or watermarks are important recommendations, enforcing their use could prove difficult to achieve, especially in light of concerns over malicious non-state or state actors unwilling to comply with technical safeguards. Computer experts also argue that there seems to be a high risk, “too high a risk”, of misfire or targeting of an innocent party due to misattribution if defensive measures are deployed with automated retaliation capability.[44] 44 Countries have now expressed concern over the challenges posed by fully autonomous lethal weapons since the May 2013 Human Rights Council.[45] A decision was also adopted in November 2013 by states party to the Convention on Conventional Weapons (CCW) to hold inaugural international discussions in May 2014 on how to address some of these challenges, including assurance of meaningful human control over targeting decisions and the use of violent force. The Campaign to Stop Killer Robots,[46] a new global campaign comprising 45 non-governmental organisations in 22

[40] Guarino, Autonomous Intelligent Agents.

[41] Tyugu, Command and Control of Cyber Weapons.

[42] Guarino, Autonomous Intelligent Agents.

[43] Guarino, Autonomous Intelligent Agents.

[44] Dmitri Alperovitch, “Towards Establishment of Cyberspace Deterrence Strategy”, *3rd International Conference on Cyber Conflict*, Tallinn, 2011.

[45] Campaign to Stop Killer Robots, http://www.stopkillerrobots.org/2013/11/ccwmandate/.

[46] Stuart Hughes, “Campaigners call for international ban on ‘killer robots’”, http://www.bbc.co.uk/news/uk-22250664, 23 April 2013.

countries, also recommends that states develop national policies and that negotiations should begin on a treaty to ban these weapons.

Though developing national policies is a good starting point, and while national legislation and international treaties are important, the regulating of such future developments could be difficult. An outright ban could be close to impossible to enforce while pursuing agreement by way of an international treaty could also raise its own particular difficulties. Further, not only can regulations be untimely in the context of rapid technological development but the controlling of these technological developments could be difficult, even where controls are put in place. It is safe to conclude that if a tool can be developed, it is more than likely that it will be developed. Cyber capabilities in particular are inherently difficult to prevent from being created and such regulatory solutions might not deter malicious actors. In addition, non-state actors will not necessarily feel morally or legally bound in the same way and state actors may not always play by the same "version of the rules".[47] A combination of technical and legal safeguards is required but further research is still needed to examine whether more could be done, while also ensuring that innovation is not suppressed disproportionately.

Public perception and acceptance of these technologies also requires further active attention as soon as possible since it could significantly impact the future uses of these technologies (although this might not be the case in every country). For instance, the public's understanding of AI and autonomous systems could fuel misconceptions about sci-fi doomsday scenarios. Alternatively, reports consider that concern over casualties could make these systems seem more attractive,[48] even if cyberwarfare could also lead to violent and destructive consequences.[49] Recently for example, the Campaign to Stop Killer Robots was created so as to demand a pre-emptive ban on the development, production and use of weapons capable of attacking targets without human intervention, in other words fully autonomous "human-out-of-the-loop systems". And in light of the recent privacy and security scandals, a number of advanced technologies developed by the public sector have already begun to be shelved in some countries over policy-related concerns.

To some extent, the public debate has already begun to kick off with a number of TED (Technology, Entertainment, Design) talks and sensational reporting. However, further widespread public discourse should be held and the public should be responsibly informed as soon as possible so that decisions may be made on many of these issues in an educated manner. Such proactive initiatives might go some way to ensure misperceptions are actively prevented before misunderstandings and possible negative perceptions become the norm. As the Director of DARPA (Defense Advanced Research Projects Agency) in the United States recently stated, these cutting-edge technologies will continue to be pushed and developed at an increasingly fast pace and society needs to begin making some important decisions about these questions.[50]

Where the public sector might be restrained from using some tools, it is still probable that they will eventually make their way into the commercial sector, if not already developed by the

[47] Under Secretary of Defense, *Resilient Military Systems*.

[48] DCDC, *Global Strategic Trends*.

[49] Mariarosaria Taddeo, "An Analysis For a Just Cyber Warfare", *4th International Conference on Cyber Conflict*, Tallinn, 2012.

[50] American Forces Press Services, "Director: DARPA Focuses on Technology for National Security", 15 November 2013.

private sector itself. It is therefore unclear whether the public or private sector will drive these technological developments in future. Defence reports suggest that financial constraints and reduced military budgets might further impede the public sector for instance, with particular financial strain from large weapons programmes,[51] in which case the perceived cost efficient aspects of these future technologies could make them more appealing. Further, the public sector does not always, and may not in future, match the speed of innovation in IT in the private sector. Defence officials explain that defence departments might have unique IT needs for example[52] and traditional ways of acquiring technologies which in some cases take many years. In the U.S. for instance this has traditionally taken close to seven years as compared to the development of the iphone which took two years. Lastly, while commercial off-the-shelf products could allow cost savings, security and supply problems might arise that endanger the security and availability of systems.[53]

For now, comprehensive guidelines that examine these concerns and policy gaps could greatly assist policy-makers by providing an informative and independent high-level analysis. A concrete examination of all the various scenarios that could possibly arise should be produced so that plans and strategies can be formulated now to prepare for all future expected as well as far-fetched outcomes. Care should also be taken to ensure that the policy formation process is informed by a deep technical understanding of how these technologies function, and that the public are engaged as much as possible as significant stakeholders. Currently, there is a wide gap that needs to be narrowed between the levels of understanding of those working in this field vis-à-vis policy-makers and the general public.

7. CONCLUSION

In summary, employing AI techniques and intelligent solutions for current as well as future cyber-related challenges, and in particular for active cyber defence, raises a number of significant technical questions and policy-related concerns. While advanced solutions are considered necessary, there is still much technical and policy-related uncertainty surrounding the future consequences of these tools, especially fully autonomous intelligent agents and possible disruptive technologies that combine AI with other disciplines. Several policy implications are highlighted that could perhaps arise such as legal uncertainty, ideological and ethical concerns, public perception problems, public-private sector ramifications, and economic issues. These policy gaps require even further examination and forward-looking solutions should be developed presently in order to anticipate difficulties that might arise in light of expected rapid developments in this field.

[51] DCDC, *Global Strategic Trends*.

[52] Lynn, 2010 Cyberspace Symposium.

[53] Koch & Rodosek, “The Role of COTS Products for High Security Systems”, *4th International Conference on Cyber Conflict*, Tallinn, 2012.