2013 5th International Conference on Cyber Conflict
K. Podins, J. Stinissen, M. Maybaum (Eds.)
2013 © NATO CCD COE Publications, Tallinn

Autonomous Intelligent Agents in Cyber Offence

Alessandro Guarino
StudioAG
Vicenza, Italy
a.guarino@studioag.eu

Abstract: Applications of artificial intelligence in cyber warfare have been mainly postulated and studied in the context of defensive operations. This paper provides an introductory overview of the use of autonomous agents in offensive cyber warfare, drawing on available open literature. The study supplies an introduction to the taxonomy and science underlying intelligent agents and their strengths and weaknesses: the technological elements that autonomous agents should include are explained, as well as the economics involved. The paper also aims to explore possible legal implications of the use of autonomous agents and shows how they could fit into the legal context of warfare. The conclusion of the study is that the use of AI technologies will be an important part of cyber offensive operations, from both the technological and the economical aspects; however, the legal and doctrinal landscape is still uncertain and proper frameworks are still to be developed.

Keywords: *artificial intelligence, autonomous agents, cyber warfare, cyber attack, international law*

1. INTRODUCTION

Cyber warfare is more and more a moving target: developments in the field are rapid, especially in the technological arena, and artificial intelligence (AI) techniques are more and more at the heart of applications. The concept of agents has been known for some time and software with some agent characteristics is already present and deployed, but in the near future we will probably see the birth of true autonomous agents, which will be a new breed entirely. This paper will propose a detailed definition of their capabilities and of what it will take to be considered truly an autonomous intelligent agent, as well as describing how their advent could fit into the international law of war. We are conscious that this subject can be considered highly speculative at the moment, and indeed it is; but until now, the discussion on international regulation of cyber warfare has been virtually nonexistent outside specialist circles and a debate on possible updates of international law to accommodate these new developments is sorely needed, especially considering that the very probable deployment of AI techniques could trigger an escalation in their use. This paper focuses on offensive activities in cyber warfare, i.e. 'cyber offence', including both Computer Network Attack (CNA) and Computer Network Exploitation (CNE), as defined in the U.S. Joint Publication 3-13. So, cyber offence includes 'actions taken via computer networks to disrupt, deny, degrade, or destroy the information within computers and computer networks and/or the computers/networks themselves' and 'actions and intelligence collection via computer networks that exploit data gathered from target or enemy information systems or networks'.

2. AUTONOMOUS AGENTS

A. CHARACTERISTICS

In the field of AI there exists a traditional division between two concepts: 'strong' and 'weak' AI: strong AI aims at creating nothing short of what the name implies, namely an intelligent being of a different species than humans but at least as capable, while weak AI assigns itself the more limited target of replicating single feats of intelligent behaviour, not surpassing human intelligence, e.g. computer vision systems, game-playing software, etc.

We leave aside the philosophical debate that the 'weak vs. strong AI' discussion entails, which is deeply interesting but alien to the aims of this paper. We concern ourselves here with 'autonomous (intelligent) agents': whatever their exact definition, surely they do not (yet) belong to the realm of strong AI and are rather applications

of various technologies mimicking natural intelligence. Autonomous intelligent agents can be purely software, or integrated into a physical system ('robots') – the difference lies mainly in the environment in which the agent operates: while purely software agents live in what we call 'cyberspace', robots can sense and interact with the same physical environment that we live in. The environment makes a lot of difference for autonomous agents, as we shall see, but the similarities between software agents and robots are relevant, given that even in a robot the embedded software – or firmware – is at the heart of its behaviour and capabilities.

Autonomous agents are a special kind of computer program, but what makes them special? Basically, every computer program is autonomous in a way – this is what computers are for, after all – so we need to develop a useful definition, especially for the field we are concerned with, cyber warfare. Franklin and Graesser [1] have given a convincing definition of agents and the ways in which they differ from other software:

> *An autonomous agent is a system 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.*

This formulation contains some very important points:

1. An agent is strictly associated with its environment: an autonomous agent outside the environment it was designed for can be useless, or not an agent at all.
2. An agent interacts with the environment, via appropriate sensors providing input from it and appropriate actuators allowing the agent to act and influence that environment.
3. An autonomous agent acts towards a goal, it has an 'agenda' in the words of Franklin and Graesser. In particular, an autonomous agent developed for warfare operations is assigned a target.
4. The activities of a truly autonomous agent are sustained 'over time', so it must have a continuity of action.

Some characteristics are probably missing here, which are required to round up our definition, even if it can be argued they are implicit in the above formulation. First, an autonomous agent needs to possess an internal representation of its environment, or what is called a 'belief state'. In their standard textbook *Artificial Intelligence – A modern approach,*[2] Stuart Russell and Peter Norvig introduce a classification of agents that ranges from simple reflex agents, where there is no internal model of the environment, through model-based agents and goal-based agents, ending with utility-based agents. In utility-based agents we find an internal representation of

the environment, a prediction of what it will be like and how near the goal it will be after the agent's actions as well as a measure of utility, i.e. a way of expressing preferences among various states of the environment (or the agent's 'world').

The utility function can be considered a measure of the performance of an agent and is the base for Russell and Norvig's definition of 'rational agent': an agent can be called 'rational' if – given the input from the environment and its internal knowledge – it will select the action or actions expected to maximize its performance measure, or utility function. The above is a paraphrase from the original.[3]

Finally, we should stress that, for an agent to be truly intelligent, the internal knowledge and the utility function itself should change over time responding to the experience acquired, or, in other words, an autonomous agent should learn from experience. This can even include modifications of the goal – the target – and can have deep ramifications for autonomous agents employed in cyber offence operations.

We should therefore round up the previous four points with two more, to achieve a complete definition of a truly autonomous intelligent agent:

5. An autonomous agent should possess an adequate internal model of its environment, including its goal – expressed possibly in terms of world-states – together with some kind of performance measure or utility function that expresses its preferences.
6. An agent must possess the capability to learn new knowledge and the possibility to modify over time its model of the world and possibly also its goals and preferences.

B. *TAXONOMY*

Artificial autonomous agents can first of all be divided into the two already mentioned classes of robotic and computational agents. Within the class of computational agents, i.e. purely software agents or 'softbots', we propose a further classification based on two coordinates that can usefully be incorporated into policies and strategic and tactical cyber operations doctrines.

Based on their role, autonomous agents can be employed in intelligence-gathering or in purely military operations: the main difference lying in the destructive nature of military operations, while usually intelligence-gathering does not cause damage to the targets and in fact tries to avoid detection in most instances. Based on architecture, autonomous agents can be either monolithic or decentralised. Monolithic agents are constituted by a single piece of software or else by strictly coordinated elements

without independent means of operation, for instance an executable file and some software libraries or data needed for it to work. Decentralised intelligent agents are systems where the intelligence is distributed among many simpler components, all similar or very similar, acting in concert, in a way similar to the artificial life 'flocks' developed by Craig Reynolds.[4] A decentralised agent can arguably be more resilient to disabling efforts and counterattacks: for instance, a botnet made of agents instead of conventional malware software would not present a central point of control that could be disabled.

Truly autonomous agents used in cyber warfare are not known at this time, at least in unclassified sources. It can be argued that, at the present stage, we are on the verge of seeing actual agents deployed, but for now the most advanced cyber weapon known – Stuxnet – falls short in many of the attributes we postulated for an autonomous (computational) agent:

- It does not possess any representation of its environment, for instance the topology of the network it is running on.
- Its action does not present a continuity in time.
- It does not have any learning capability.
- It does not perform an autonomous target evaluation or selection. It is true that it is capable of selecting systems according to a set of targeting criteria, but the set is fixed at programming time and not subject to expansion or modifications, exactly because no learning is involved.

C. *THE ENVIRONMENT OF SOFTBOTS: 'CYBERSPACE'*

Physical autonomous agents, or 'robots', operate in the normal physical environment, while computational agents operate in a unique environment, what is commonly called 'cyberspace'.

A valid definition of cyberspace is given in the White House Cyberspace Policy Review published in 2011. According to this document, cyberspace is 'the interdependent network of information technology infrastructures, and includes the Internet, telecommunications networks, computer systems and embedded processors and controllers in critical industries'.[5] Cyberspace is unique as an environment in many ways, but most of all because it is man-made, and not by a simple subject: it is constructed, maintained and operated by a plurality of stakeholders, public and private and with a multitude of somewhat conflicting interests and incentives.[6]

Following Russell and Norvig's characterisation of environments, we can list other peculiarities of cyberspace:

- Cyberspace is a partially observable environment, meaning that agents can have, at the best of times, knowledge of only a tiny fraction of it.
- It is a deterministic and discrete world, but of enormous complexity.
- For agents engaged in cyber warfare the environment is obviously adversarial, including enemy operators, enemy and foreign agents, and conventional security software like firewalls, Intrusion Detection Systems (IDSs) and Intrusion Prevention Systems (IPSs).
- Cyberspace is dynamic, meaning that the agent's environment can and will change during its operation.

3. CYBER ATTACK SCENARIOS FOR AUTONOMOUS AGENTS

Most of the public debate about cyber warfare policy and strategy in recent years, including academic production, has concentrated on defence. There are valid reasons for that, including: the obvious secrecy shrouding offensive tools and procedures; political reasons; access to information; and a bias towards defence in the West, always shy of appearing as the aggressor. In fact, in cyber warfare, not only does offence have a place in strategy, but an offensive stance – or at least an active defence – is probably preferable. Also, an offensive stance is easier to adopt if the perceived costs – in terms of casualties but also monetary and political costs – are very low compared to other forms of warfare.[7]

A. RECONNAISSANCE

All offensive operations begin with reconnaissance, and this first phase of a cyber attack will arguably provide an ideal arena for the deployment of autonomous agents in the near future, at least in two directions: automatic discovery of technical vulnerabilities in target systems or networks and, on another level, gathering of intelligence about them, for instance structure and topology and details of operating systems and applications, up to user details and access credentials.

The discovery of vulnerabilities in the target network, and the development of practical means of leveraging them (called '*exploits*'), are necessities; presently they are manually developed by skilled personnel or acquired on the market. A software autonomous agent will automatically reconnoitre the target, individuate vulnerabilities and develop means of exploiting them: while full automation of exploit development is still not widely available,[8] we can outline some scenarios for the use of agents incorporating such a capability:

- Software agents instructed to target a network. In this case, the operation will proceed beyond the information-gathering phase into actual infiltration; moreover, a target is already selected. The agents, however, will have no need for fixed, pre-programmed methods of penetrating the system, but will analyse the target, select autonomously the points of vulnerability and develop means to use them. The agent will then proceed to actual infiltration of the target system and – if it is operating as part of a decentralised agent – will share the information gathered with the other agents.
- Purely information-gathering agents. In these kinds of operation, the main objective of the agent is the acquisition of information, which will then be sent back to the command and control structure where it will be processed. Technical information about vulnerabilities present and exploits can be entered into a database that can be used by multiple operations.

B. INFILTRATION AND BEYOND

Autonomous agents, as defined here, will be able to 'remember' the information gathered during reconnaissance and use it to plan their infiltration path. One of the possible methods makes use of 'trees' – mathematical structures commonly used to represent AI problems – to model the possible alternatives in a cyber attack. [9] Future agents will conceivably be able to build *ad hoc* tree representations of possible infiltration routes on the fly, as opposed to manually, and apply techniques – some of them already very well established – to plan and execute the infiltration. Internal representation of the environment and possible threats from defenders will make it possible for autonomous agents to be much more 'persistent' than advanced persistent threats (APTs) known today, by allowing them to prevent and react to countermeasures: for agents tasked with intelligence-gathering, this will mean more time to do so, and agents tasked with disruption will enjoy much more flexibility in selecting specific targets (applications or systems) and means of attacking them. The selection of, for example, specific databases or documents to retrieve once the agent gains access would be achieved through AI techniques that can extract and process information even from unstructured data.

C. SWARMS

Decentralised agents, according to the taxonomy presented above, are sets of cooperating autonomous agents, that can form a whole new kind of botnet, where there is no need for a centralised command and control and individual agents can cooperate, amongst other things, by sharing information. A botnet of this kind would be very difficult to disable, because each single agent would be separately individuated and sanitised.

D. COMMAND AND CONTROL

From an operational standpoint, the difficult problem of command and control of autonomous agents should be stressed. On the one hand, the agent's goal and targets should be pre-programmed and precisely stated in order to facilitate the agent's task and stay as much as possible within legality. On the other hand, it will be tempting to leave free rein to an extent to agents, for instance regarding targets of opportunities. The above concerns the initial phase of developing and deploying an autonomous agent, but it should also be decided to what extent the agent could communicate with its 'base' during its mission, and if that communication should be monodirectional – intelligence originating from the agent, for instance – or bidirectional, i.e. if the command and control structure could issue 'orders' and instructions, including target selection and even self-destruct commands. It is obvious that a whole doctrine including detailed tactics, techniques and procedures (TTPs) for intelligent autonomous agents will have to be developed in order to be ready to integrate these new tools into a state's arsenal.

4. THE LEGAL LANDSCAPE

There has been general discussion on the application of international law to the case of cyber warfare, but a consensus on precise terms has not yet been reached: the possibility in the very near future of the emergence of true autonomous agents pushes the debate still further and shows even more clearly the limits of existing international law in the realm of cyber warfare. While it may be considered far-fetched at the least, and bordering on science fiction, to be discussing right now the legal implications of the use of intelligent autonomous agents as cyber weapons, information technology has a history of preceding the law by far and maybe it is not wrong to engage in such a debate earlier than usual.

The international body of law governing warfare basically consists of *jus ad bellum*, which regulates the resort to force by states; the International Humanitarian Law or *jus in bello*, which concerns itself with the actual conduct of armed conflicts; and the law of neutrality.

A. JUS AD BELLUM

The main source of the international *jus ad bellum* is the Charter of the United Nations, signed in San Francisco in 1945, and its successive amendments. Article 2(4) of the Charter concerns the use of force by Members: 'All Members shall refrain in their international relations from the threat or use of force against the territorial integrity or political independence of any state [...]'.

It is commonly accepted that this article of the Charter applies to cyber warfare too, the effects of which are 'comparable to those likely to result from kinetic, chemical, biological or nuclear weapons',[10] and in this regard autonomous agents are no different from any other tool or cyber weapon employed. Where the advent of true autonomous agents could really require new interpretations or new formulation is in the question of agency, i.e. 'the attributability of individual conduct to a state'.[11] An autonomous agent will act – up to a point – independently from its developers, at least as regards the details of its actions: we should ask ourselves if the notion of an individual acting as a 'state agent' should be extended to autonomous agents. The notion of states and their sovereignty is central to the Charter and all international laws of war, even if, post-9/11, this is somewhat less true and non-national actors have been accepted as, for instance, capable of waging war, as in the case of terrorist networks. Theoretically, a true autonomous agent could exceed its assigned tasks and engage in what could legally be defined as 'use of force': in this case, should the nation state behind the agent's creation be deemed responsible? The problem of attack attribution, so important for cyber warfare in general ,is even more problematic for attacks realised by an autonomous agent, if only for the fact that its creators themselves possibly would not have known in advance the precise technique employed, or even the precise system targeted, because of autonomous decisions taken by the agent during its operations. In other words, command and control of a true autonomous agent, especially a purely computational one, can be hard to achieve and would have to translate chiefly in precise specifications of the agent's target and objectives – the goals – or, in military terms, in precise briefings before any mission.

Another question that should form part of the debate on the legal aspects of autonomous agents in cyber warfare is whether they can be considered *per se as* 'state agents'. Again, discussion of a similar concept can seem far-fetched at this time, but we are not far from the deployment of real agents and policy-makers should be made aware of the implications. A true intelligent autonomous agent, as defined above, would possess the capability to make decisions based on its belief state of the moment and its assigned tasks, so it is reasonable to consider it a 'state agent' in a legal sense. If so, it seems reasonable to argue that a way should be found to distinguish whether a software agent is to be considered civilian or military, in a technical and a legal sense. In the case of the postulated physical agents, it is easy to assume that – as in the case of remotely controlled 'drones' – they would sport national identification marks, but what about software bots? If ever an international treaty on cyber warfare is signed, it should contain provisions for the identification of autonomous agents, maybe through mandatory signatures or watermarks embedded in their code.

The United Nations Charter does not clearly forbid the use of force in any case, but implicitly admits it in the case of self-defence, in its Article 51: 'Nothing […]

shall impair the inherent right of individual or collective self-defence if an armed attack occurs against a Member […]'. While, customarily, only attacks by national actors were deemed to be covered by the provisions of this formulation, after the events of September 11, 2001 a new concept emerged, heralded obviously by the US, whereby a nation-state is allowed, in self-defence, also to use force against non-state actors, such as the Al-Qaeda network. In order to be lawful, however, the use of force in self-defence should be governed by some principles, primarily necessity and proportionality. 'Proportionality' in this instance means that the level of force used in self-defence should be no more than what is necessary to repel the threat. Again, in the case of a true autonomous agent, if used as a weapon in self-defence, care should be taken in the command and control function to clearly state the agent's targets and build in appropriate safeguards.

Concerning the individuation of true 'armed attacks', as defined in Art. 51 of the U.N. Charter, and in particular their distinction from lesser instances of use of force ('border incidents'),[12] what is relevant is to determine the actual intent of the state operating the autonomous agent. In this case, the theme of independent behaviour of agents returns. True autonomous agents certainly offer military leaders the broadest possible extent of plausible deniability and it would be difficult to make all actions by an agent the responsibility of its creators.

B. *JUS IN BELLO (INTERNATIONAL HUMANITARIAN LAW)*

This body of law strictly governs actions and activities that happen during actual armed conflicts. Its main sources are the Geneva Conventions of 1949, the Additional Protocols of 1977 and the much earlier Hague Convention of 1907. Together with this written corpus, there is also a rich tradition of customary humanitarian law dating back at least to the Roman *jus gentium*. The dates of the applicable conventions should be enough to emphasise how problematic it is to reconcile the realities of cyber warfare with this established legal landscape, even more so because up to this point we have seen very few instances of actual cyber warfare and none involving a true autonomous agent, whether in international armed conflicts or non-international ones. The recently published *Tallinn Manual*[13] provides a much-needed guide on how international law applies to cyber warfare, even if its scope extends – by the authors' choice – only to existing law (*lex lata*) and how it is applicable to fully-fledged conflicts.

For a cyber operation to be considered an attack under international humanitarian law (IHL), it needs to occur in the context of an armed conflict (hostilities), or, in other words, to have a nexus to a conflict. In the existing law, only in this case is IHL applicable. As the definition implies, IHL is mainly concerned with the protection from violence that should be guaranteed to entities not involved in the

conflict, be they persons or physical assets. So military actions, to be lawful, should avoid – or at least minimise – what can be defined as 'collateral damage': attacks must not be indiscriminate and targets must be carefully selected. Also, the means used (weapons and tactics) should aim to avoid, as far as possible, unnecessary victims and damage to civilian assets. This principle goes under the name of 'proportionality', a somewhat different concept from the principle of the same name present in *jus ad bellum*. Precise targeting of an autonomous agent should not be difficult for computational agents, where, for instance, it can be assured either by protocol parameters such as addresses defining the boundaries of a network or by profiling beforehand what constitutes the agent's target in terms of technical characteristics, application discovered, or types of data.

As in the case of drones, where many have expressed concerns over a so-called 'PlayStation syndrome', the remoteness of the command and control operators could result in less restraint in attack operations. For cyber autonomous agents, a 'remoteness' in time is present, in addition to a distance in space similar to that for the pilots of unmanned aircraft systems, and appropriate doctrines should be developed for the planning, command and control of cyber operations involving autonomous agents. If some form of built-in identification mark is introduced for cyber weapons, this should infuse more responsibility in their planning and operation.

5. CONCLUSION

While, currently, true autonomous agents probably do not yet exist, the technological preconditions for their development are in place and active research is ongoing: it is useful, therefore, to reflect beforehand on the all-round implications of their use in warfare, considering both the technical foundations and the legal implications. The taxonomy proposed here can be considered as a basis for future works, including the place of such agents in doctrine. Agents relying on AI techniques and operating independently would be considered a force enhancer and would make offensive operations even more attractive; anonymity would be enhanced by the independent way in which such agents would operate, conducting attacks without supervision or contact with a command and control structure and for a prolonged period of time; the amount of information needed to launch an attack would be far less than with conventional cyber weapons, because the agent itself would develop its own intelligence, for instance analysing vulnerabilities and developing exploits for them. Also, an autonomous agent, especially if decentralised, would be much more resilient and able to repel active measures deployed to counter it, while conventional malware is somewhat more fragile in this regard.

If – or when – actual autonomous agents are developed and deployed, existing international law will be truly strained and will need adjustments, for instance regarding the possibility of considering a software agent as a 'state agent' as conceived in international law. Work on cooperation and agreements is necessary in order to avoid something similar to an arms race, mainly because the capabilities postulated here for autonomous agents will render cyber offence activities even more attractive than they already are.

Further research is needed both on the technical and the legal side. The author is working on a possible implementation of an autonomous agent using AI techniques and also on further developing the modes in which such agents could be deployed. On the legal side, more work will need to be done on how AI agents would fit into contexts other than a fully-fledged armed conflict, for instance in peacekeeping operations.

REFERENCES

[1] S. Franklin and A. Graesser. 'Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents', in *Proceedings of the Third International Workshop on Agent Theories*, Springer-Verlag, 1996.

[2] S. Russell and P. Norvig. *Artificial Intelligence: a modern approach* 3rd edition, Pearson, 2010.

[3] S. Russell and P. Norvig, op.cit., page 37.

[4] C. Reynolds. 'Flocks, Herds, and Schools: A Distributed Behavioral Model', in *Computer Graphics*, 21(4) (SIGGRAPH '87 Conference Proceedings)' pages 25-34.

[5] White House, *National Security Presidential Directive 54/Homeland Security Presidential Directive 23*.

[6] N. Melzer. *Cyber Warfare and International Law* - U.N. UNIDIR 2011.

[7] R. Anderson, *Why Information Security is Hard – An Economic Perspective*, University of Cambridge 2001.

[8] J. DeMott, R. Enbody and W. Punch, 'Towards an Automatic Exploit Pipeline', in *Internet Technology and Secured Transactions (ICITST)*, 2011.

[9] A. Moore, R. Ellison and R. Linger, *Attack Modeling for Information Security and Survivability* - Carnegie Mellon Software Engineering Institute 2011.

[10] N. Melzer, op.cit.

[11] N. Melzer, op.cit.

[12] N. Melzer, op.cit.

[13] M. Schmitt (ed.), *Tallinn Manual on the International Law Applicable to Cyber Warfare* - Cambridge University Press 2013.