Internal Architecture for Software Autonomous Intelligent Agents

Alessandro Guarino

Introduction

This position paper is part of an ongoing research into the role of Autonomous Intelligent Agents (AIAs) in cyber conflict and cyber warfare, started with a definition proposed by the author in 2013[i]. Among future research areas mentioned in that paper were internal architecture and information exchange.

The Six Traits of AIAs

Autonomous Intelligent Agents can be classified along two variables: purely software vs. physical autonomous agents (e.g. UCAVs) and monolithic vs. swarm architecture. This position paper deals only with software agents.

True AIAs present six characteristic traits:

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. Franklin and Graesser[ii] have given a convincing definition of agents and the ways in which they differ from other software. The first four point in our definition draw from their definition.
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, or, in other words, it has an 'agenda'. 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
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.

As far as the unclassified world goes, no truly Autonomous and Intelligent agents are known, fulfilling all the criteria.

What is missing

An AIA should be able to sense and map its environment in an autonomous way. In cyberspace this translates to the ability to map the network in which it's operating, recognising among other things its targets or way to propagate towards them. More in detail, it will need to probe for vulnerabilities and resources, avoiding detection. It should be able to build an internal representation of the environment and use it to map a path towards the mission objective it has been assigned (offensive or defensive). On the "actuator" side, it should be able to acquire information, if its part of the brief, modify existing software and data, including the ability to develop exploits on the fly. Of course the agent will need to maintain an internal "image" of the network and its current positon in it and plan how to reach the target(s) assigned. Even from this brief presentation another critical point is apparent: a sizable quantity of information have to be maintained, in addition to the agent software itself. Appropriate utility functions will have to defined to allow AIAs to operate, possibly leveraging Machine Learning models and reinforcement learning could be used for planning purposes.

True AIAs will have no rigid programming and targeting, like even the most advanced cyber weapons seen in the wild still present, but will have to be given something akin to "mission objectives" before deployment. A compact and reusable way to accomplish this have to be developed, to be used not only before the mission begins but also during deployment. An autonomous agent - for practical and legal reasons - will not be deployed without any means of controlling it. This introduces the problem of communication between the agent and its "controllers": on one way it will have to report back, transmit or exfiltrate information and data acquired; on the other it will have the capability to receive updated intelligence or a new agenda (or a self destruct command). This by itself is a complex problem.

Conclusions

Artificial Intelligence, and most probably, the sub-field of Machine Learning, provides developers with tools to build a true autonomous agent, able to fullfil the six traits we explained above in the real world. Number six in particular calls for the ability to modify autonomously the targeting (in accordance with the brief of course) and this can be achieved only the agent is able to make changes to its very structure and world representation. A lot of ground still has tobe covered however, as integrating all the pieces is proving more difficult than expected. Autonomous agents moreover should have a robust, reliable, and controllable behaviour, as the implications from the standpoint of international law are quite complex. Physical autonomous weapon systems are in the spotlight but the potential (for stealth operations against critical infrastructure for instance) of software AIAs

---

i A. Guarino, "Autonomous Intelligent Agents in Cyber Offence", in “5th International Conference on Cyber Conflict – Proceedings”, K. Podins – J. Stinissen – M. Maybaum (eds.), IEEE 2013 – ISBN 978-9949-9211.4.0

ii 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.