Towards a framework for policy development in cybersecurity

Security and privacy considerations in autonomous agents

V.1.0

DECEMBER 2018

www.enisa.europa.eu European Union Agency For Network and Information Security

About ENISA

The European Union Agency for Network and Information Security (ENISA) is a centre of network and information security expertise for the EU, its member states, the private sector and EU citizens. ENISA works with these groups to develop advice and recommendations on good practice in information security. It assists member states in implementing relevant EU legislation and works to improve the resilience of Europe's critical information infrastructure and networks. ENISA seeks to enhance existing expertise in member states by supporting the development of cross-border communities committed to improving network and information security throughout the EU. More information about ENISA and its work can be found at www.enisa.europa.eu.

Contributors

Kalloniatis Christos (University of the Aegean), Costas Lambrinoudakis (University of Piraeus), Konstantinos Kotis (University of Piraeus), Prokopios Drogkaris (ENISA)

Editors

Prokopios Drogkaris (ENISA), Athena Bourka (ENISA)

Contact

For queries in relation to this paper, please use isdp@enisa.europa.eu
For media enquires about this paper, please use press@enisa.europa.eu.

Acknowledgements

We would like to thank Spyros Kokolakis (University of the Aegean), Stefanos Gritzalis (University of the Aegean) and Apostolos Malatras (ENISA) for their contributions during the preparation of this document.

Legal notice
Notice must be taken that this publication represents the views and interpretations of ENISA, unless stated otherwise. This publication should not be construed to be a legal action of ENISA or the ENISA bodies unless adopted pursuant to the Regulation (EU) No 526/2013. This publication does not necessarily represent state-of the-art and ENISA may update it from time to time.

Third-party sources are quoted as appropriate. ENISA is not responsible for the content of the external sources including external websites referenced in this publication.

This publication is intended for information purposes only. It must be accessible free of charge. Neither ENISA nor any person acting on its behalf is responsible for the use that might be made of the information contained in this publication.

Copyright Notice
© European Union Agency for Network and Information Security (ENISA), 2018
Reproduction is authorised provided the source is acknowledged.

ISBN 978-92-9204-282-0, DOI 10.2824/453043

Table of Contents

Executive Summary

Over the last years, Artificial Intelligence (AI) has rapidly moved beyond the realms of research and academia to enter the commercial mainstream, with innovative autonomous agents utilizing AI and transforming how we access and leverage information. Autonomous agents are characterized by diversity, with applications varying from digital assistants residing in our smartphones to autonomous robots supporting the supply chain. Depending on the level of autonomy and the context of operations, security and privacy considerations may vary. One of the key aspects in autonomous systems is the data collected, mainly for supporting the demanding functionality in a qualitative and timely manner. However, due to the abundance of data processed, which may also include personal data, in addition to the relying on third party providers, introduces a number of security and privacy considerations. The current study highlights a number of relevant considerations, such as unauthorized autonomous systems, hijacking and misuse transparency and accountability, pervasiveness, retention and opacity of processing. During this analysis, a set of recommendations for shaping future EU policy initiative was deducted which includes:

- European Commission, relevant EU Bodies and relevant public and private sector stakeholders should further promote and support the adoption of security and privacy by design principles as a pre-requisite during the inception, design and implementation of autonomous agents and systems.
- European Commission, relevant EU Bodies, public and private sector stakeholders should foster a collaborative approach on identification and exchange of best practices. Gradually such initiatives should put forward sets of baseline security requirements which can then be transposed to widely accepted technical specifications and standards.
- European Commission, relevant EU Bodies, public and private sector stakeholders should further endorse and support existing initiatives on promotion and protection of human rights through establishing appropriate ethical parameters.

1. Introduction

1.1 Background

The rapid adoption of digital technologies in recent years has enabled many opportunities and allowed for real-time adjustments and decisions to be made, based on the amount of relevant available data and processing capabilities. Such an environment presents unlimited opportunities for innovation and interaction, but bring also along a number of challenges which should be addressed by future policy frameworks at EU level. Artificial intelligence (AI) and increasingly complex algorithms currently offer great potential and possibilities for a diverse range of application areas which influence our everyday lives more than ever before. In computer science, Artificial Intelligence (AI) is a way of making software think intelligently, in a similar manner to the way humans think (Garbhe, 2017). These tasks can be cognitively challenging and hence an observer might mistake the behaviour of the machine for the behaviour of an intelligent human. The field of AI however, attempts not only to understand an intelligent entity, and the way it perceives, understands, predicts, and manipulates a world far larger and more complicated than itself, but also attempts to build one (Russell, Norvig, & Davis, 2010). AI attempts to understand the way the human brain thinks and how humans learn, decide and work, while trying to solve a particular problem, and then make use of this understanding as a basis for developing intelligent software and systems.

1.2 Initiatives at EU level

The European Commission identified, in its mid-term review of the Digital Single Market strategy[1], the importance of building on Europe's scientific and industrial strengths, as well as on its innovative start-ups, to be in a leading position in the development of AI technologies, platforms, and applications. In mid-2018, the European Commission, through COM(2018) 237[2], highlighted the need for the EU to become a leader in the AI revolution while acknowledging that "*Like the steam engine or electricity in the past, AI is transforming our world, our society and our industry*"[1]. The Communication sets out a European initiative on AI, which aims to:

- Boost the EU's technological and industrial capacity and AI uptake across the economy, both by the private and public sectors, including investments in research and innovation and better access to data.
- Prepare for socio-economic changes brought about by AI by encouraging the modernisation of education and training systems, nurturing talent, anticipating changes in the labour market, supporting labour market transitions and adaptation of social protection systems.
- Ensure an appropriate ethical and legal framework**,** moving a step forward on ensuring legal clarity in AI-based applications.

Lastly, the European Commission recently appointed the members to the High-Level Expert Group on Artificial Intelligence (AI HLEG) which seeks to support the implementation of the European strategy on Artificial Intelligence[3].

[1] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2017:228:FIN
[2] https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe
[3] https://ec.europa.eu/digital-single-market/en/artificial-intelligence

1.3 Scope – Objectives

Within the scope of the ENISA 2018 Annual Programming Document[4] and towards supporting a framework for policy development in the area of cybersecurity, the Agency undertook the current study in an attempt to highlight the main security and privacy considerations that arise from the use of autonomous agents in today's modern society. The objective of the study is on the one hand to provide an overview of representative application domains and on the other hand to contribute and complement relevant initiatives at EU level while also providing relevant insights, both for security and privacy, for future EU policy shaping initiatives.

As a first step towards sketching a point of reference for policy-makers tackling the issue of cybersecurity and establighing a relevant framework for policy development, emerging technologies and new application areas must be considered, embracing the new opportunities they bring along but in a way that they respond to European needs. As also acknowledged by the EC Joint Research Centre (JRC) report titled "Artificial Intelligence: A European Perspective"[5], "the EU Member States and the European Commission are developing coordinated national and European strategies, recognising that we can only succeed together".

1.4 Structure of the study

Section 2 of the study outlines AI technology, which has been used in the context of autonomous agents, in several application domains through a representative, rather than an exhaustive, list. Section 3 discusses the main security and privacy considerations by aggregating and building up on already published ENISA studies in the area of IoT and Smart Infrastructures[6] while in Section 4 a set of recommendations for relevant stakeholders and policy makers is discussed.

[4] ENISA programming document 2018-2020, https://www.enisa.europa.eu/publications/corporate-documents/enisa-programming-document-2018-2020.
[5] Artificial Intelligence: A European Perspective, https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/artificial-intelligence-european-perspective
[6] https://www.enisa.europa.eu/topics/iot-and-smart-infrastructures

2. Artificial Intelligence and Autonomous Agents

Modern Artificial Intelligence (AI) applications are pervasive and numerous to list. Frequently, when an application/technique of AI reaches mainstream use, it is no longer considered artificial intelligence. Some of the most widely used AI applications are included in domains such as:

- healthcare (e.g., knowledge-based diagnosis, medical personal assistants for elderly, decision-support systems for cancer treatment, robotic surgery),
- medical robots for surgical purposes in cases that do not necessarily require human medical intervention,
- automotive (e.g., driverless cars, distributed multi-agent coordination of autonomous vehicles),
- finance/economics (e.g., fraud and financial crimes detection, AI based trading platforms),
- gaming (e.g., dynamic purposeful behaviour in non-player characters (NPCs), pathfinding, deep learning prediction),
- military (e.g., autonomous drones),
- security (e.g., speech/image/object/face-recognition),
- advertising (e.g., predicting the behaviour of customers),
- art (e.g., e-auctions, automated story-telling creation, digital museum guide, automated music synthesis),
- social life (e.g., digital personal assistants, smart homes).

An AI system is mainly composed of a software agent and the environment in which it operates (sense and act). Agents act in their environment, which may contain other agents too. The first two decades of the 21st century are characterized by a number of examples of autonomous technology and AI. Most existing AI systems and robots are automatic but not autonomous i.e. they do not develop and maintain their internal structure and functioning through mechanisms like self-organization, evolution, adaptation, and learning. Driverless cars and drones, robots in deep sea and space exploration, weapon systems, bots in financial trade and deep learning in medical diagnosis, are among the most prominent, but certainly not the only examples of autonomous agents. AI, especially in the form of machine learning, and the increasing availability of Big Data, are important drivers of these developments. The coexistence and combination of these digital technologies is rapidly making them more powerful, and they are applied in an increasing number of new products and services, in public and private sectors, and can have both military and civilian use. The AI used in these systems can redefine work or improve work conditions for humans and reduce the need for human contribution, input and interference during operation. It can help to assist or replace humans with smart technology in time-consuming, expensive, difficult, dirty, dull or dangerous work, and even beyond.

The typology of such systems includes combined physical/logical ones such as a driverless car/boat/aircraft; and logical bots such as a software agent e.g. Microsoft Cortana, Amazon Alexa or Apple Siri.

2.1 Autonomous agents as a business

For the time being autonomous agents are addressed in the context of a specific business case e.g. delivery of packets by a drone or a public service e.g. surveillance of an area for fire hazard by a drone. Scoping the operation and the permitted actions by an agent is of paramount importance. Often contradicting requirements need to be balanced out to produce a meaningful outcome. The risk of breach of privacy is significant therefore when hovering in inhabited areas, specific limitations to collecting data should be put

in place; by the same token such limitations should be lifted in case of surveilling a critical infrastructure, and be imposed within reason for e.g. individuals who happen to walk in the area. Suitable provisions need to also be put in place to be able to collect data in case criminals must be identified by Law Enforcement Authorities (LEAs).

2.2 Knowledge representation

Intelligent agents are software entities that carry out a set of operations on behalf of a user or another program with some degree of independence or autonomy. Doing so, they employ some knowledge or representation of the world and the users' goals or desires. Knowledge-based autonomous agents (Zhou, Liao, & Tang, 2012) are best understood as agents that have knowledge about their world and reason for their actions. They usually integrate a) a knowledge-base (KB) i.e. a set of representations of facts about the world, and b) a knowledge representation language i.e. a language whose sentences represent facts about the world. Specific operations for adding new sentences to the KB and querying what is known are required, as well as a reasoning mechanism for determining what follows from what has been specified to the knowledge base. We can build a knowledge-based agent by 'telling' it what it needs to know before it starts perceiving the world, and we can design learning mechanisms that output general knowledge about the environment given a series of perceptions. In other words, an autonomous agent is a knowledge-based agent with an integrated learning mechanism.

2.3 The notion of autonomy

Autonomy is rooted in the ability to set and abide by a self-imposed rules based system. In autonomous systems this notion translates to the ability to apply and interpret (to a certain extent) rules. This ability combines with a learning capacity that renders the outcome impossible to predict. Safeguards need to be put in place to curtail broad discretionary autonomy and render predictable the behaviour of such systems and introduce specific limits and boundaries. While meaningful human control might not be possible at all times, it is important to lay down the technical and procedural framework to render such controls possible.

2.4 Learning

An AI system needs large volumes of data to first analyze and then learn from them. In order to be able to learn, it first needs to be trained on how to do this. Generally, learning systems aim to a) recognize patterns (e.g., recognize a cat from a car), b) detect anomalies (e.g., detect fraud in expense reports), and c) predict (e.g., predict stocks). To accomplish these, the following types of learning may be used, as those are simply described in Layman's intro to AI and Neural Networks (Preetham V V, 2016):

- **Supervised Learning**. In this type of learning, there is need to know what to expect from the system in the output. The systems, during their training, are given the output needs. For instance, the English alphabet can be used by the system as a training dataset to support the recognition of English handwriting. Given a set of documents, the system must be able to predict any English handwriting included in them. During the training of the system, a few handwritten alphabets are entered into the system as input, along with the desired matching output alphabet. Once the system gains accuracy in recognizing handwritten alphabets, it can be used, for instance, in autonomous driverless cars to read signs on the highway.
- **Unsupervised Learning**. In this type of learning, the systems do not 'know' the output needs. On the other hand, they are allowed to expose the internal representation of the input. For instance, automatically categorizing images with cats in a different category/class of the recognized images of a car. However, the system cannot recognize and label the car category as a "car, vehicle" or the cat category as "cat, animal" since the semantics of the categories are not given as input in the system (unlike supervised learning, which can recognize and label).
- **Reinforcement Learning**. In this type of learning, the systems are asked to increase the payoff of an activity by selecting the most optimal action. While there may not be a predefined expected output, the

systems can look at what payoffs are voted higher and optimize actions. For instance, recommending the music that we like based on our music preferences. In autonomous agents, learning trajectory patterns of autonomous vehicles, based on real-time analysis of large volumes of data, may result to recommending the most economic trajectory from origin A to destination B, given specific weather and air-traffic conditions. Latest reports demonstrate how an autonomous car earned how to drive itself in 20 minutes using reinforcement learning[7].

2.5 Neural Networks/Deep Learning

Artificial Neural Networks (ANN) is a significant technology for the control of autonomous agents. The major goal of the research on autonomous agents is to study intelligence as the result of a system environment interaction, rather than understanding intelligence on a computational level. Autonomous agents have to continuously learn while they are behaving in their environment, without distinguishing between a learning and a performance phase. Thus, a neural network for autonomous agents should feature incremental/continuous learning (Salomon, 1996).

In an autonomous agent scenario, the agent initially starts to explore its environment and, during its lifetime, it gains more and more experiences. In such a scenario it is clear that: i) an agent determines the pattern's relevance at runtime, ii) it is not appropriate to select a set of training examples for off-line training, and iii) an agent needs a network (for example a neural one) that features incremental learning properties, since the agent acquires its experience over time.

For an autonomous agent, this means that it must update its network weights while it is operating. The network model for autonomous agents should be elastic since during its lifetime, the appropriate actions associated with a particular sensory information might change. This dynamic change can be caused by a change in the environment (perhaps due to the agent's interaction with its environment) or by a dynamical change of the electrical characteristics of the sensors. Simply put, a Neural Network is a collection of nodes (or units) which are layered as input nodes, hidden nodes and output nodes. The input nodes can be, for instance, the pixels from a vehicle image, the hidden nodes shall learn the 'knowledge weight' of the vehicle-ness of a car, and then one of the output node is selected based on the type of the vehicle (car, bus, train, drone, etc.).

2.6 Self - management

In a self-managed autonomous system, the user, instead of controlling the system directly, defines general policies and rules that the guide system's self-management process. Four types of self-management properties have been currently specified, referred to as self-star (also found as self-*, self-x, or auto-*) properties (Sanz, López, Bermejo, Chinchilla, & Conde, 2005)(Poslad, n.d.):

- **Self-configuration**: Automatic configuration of systems' components
- **Self-healing**: Automatic recognition and healing of systems' faults
- **Self-optimization**: Automatic optimization of functioning based on monitoring and control of resources with respect to the defined requirements
- **Self-protection**: Proactive identification of and protection from arbitrary attacks.

We further outline an expanded set of self-star properties as proposed by Poslad (Poslad, n.d.) as well as by Nami and Bertel (Nami & Bertels, 2007):

[7] https://www.analyticsvidhya.com/blog/2018/07/autonomous-car-learnt-drive-itself-20-minutes-using-reinforcement-learning/

- Self-regulation: Systems operate to maintain a particular level of some parameter without external control, e.g., Quality of service
- Self-learning: Systems use Machine Learning techniques to learn how to process new data/information (unsupervised learning) without external control
- Self-awareness (also called self-inspection and self-decision): Systems have knowledge of themselves, are aware of the internal components and external links in order to control and manage them
- Self-governance: Systems manage themselves without external intervention. Self-governance (or self-management) also refers to a set of self-star processes such as autonomous computing rather than a single self-star process
- Self-description (also called self-explanation or self-representation): Systems explain themselves, they are capable of being understood (by humans) without further explanation.

Autonomous agents should be able to demonstrate one or more self-x properties in order to claim their 'intelligence'.

2.7 Decision-making systems

A decision support system (DSS) is a computer-based application that collects, organizes and analyzes business data to facilitate quality business decision-making for management, operations and planning. A well-designed DSS should support the decision makers to analyse a variety of data (e.g., raw data, documents, personal knowledge from employees, etc.).

Research on DSSs mainly concerns the integration of agents in applications. Such work has been well documented, including studies that can lead to the successful integration of agents in DSS (Hess, Rees, Matheson, & Ragsdale, 1999)(Hess, Rees, & Rakes, 2008). AI techniques have been widely used in research for developing new tools for decision-making (e.g., clinical). In most of them, the computer supports users (e.g., clinicians, patients) who take the decisions. On the other hand, it is simple to turn decision support tools into autonomous agents, i.e., tools that make the decisions themselves, without external control.

For instance, the CREDO decision support platform (Fox & Khan, 2014), implementing data interpretation and decision support capabilities in a generic technology for clinical decision support and multidisciplinary patient care, builds on a cognitive model of decision making-based argumentation theory and formalized in agent specification and knowledge representation language (Fox & Khan, 2014). Such a system could be switched in an autonomous mode by properly facilitating: a) Intelligibility: a healthcare agent should be able to engage in natural and cooperative interaction with its users (clinicians, patients), b) Personalization: an agent should accommodate the users' personal goals and preferences, as long as they don't conflict with specific principles, c) Justifiability: an understandable rationale must be available for all recommendations, and particularly for automated actions, at whatever level of detail the user may reasonably require, d) Controllability: the user must be able to modify the system's assumptions and goals, and the system must be able to adapt appropriately and safely to such changes.

2.8 Recognition Technology

A recognition system integrates technology that is capable of identifying or verifying an entity or object (e.g., a person, a car, an animal, a packet, a sign, a lane, a light) from the analysis of (streaming) data that is collected by sensors (e.g., camera, microphone, laser, thermometer, pulse meter) and provide in the form of a digital file (such as an image, a video or sound file) or a stream. Recognition of traffic jams, traffic or car lights, notification or traffic signs, lanes, vehicles, pedestrians and ambient sounds are only a few examples of the recognition technology in the transportation domain for facilitating autonomy in driverless vehicles. Moreover, the integration and combination of such technology increases the performance of such complex systems in unique ways e.g., face recognition increases the safety of autonomous vehicles (Arar, 2016).

2.9 Planning

Planning in AI regards the realization of strategies or action sequences for execution of autonomous agents, thus, it is usually used in the phrase “automated planning and scheduling”. Technological approaches that provide solutions for AI planning include but are not limited to: dynamic programming, reinforcement learning and combinatorial optimization. Coordination endows rational agents with cooperation abilities to generate coherent activities and ensure a correct Multi-Agent System (MAS) behaviour. Such coordination requires an adequate plan representation (Fallah-Seghrouchni, 2005). Agents organize their activities and update their plans to cooperate and avoid conflicts. Multi-agent planning remains one of the main mechanisms for MAS coordination. It raises several interesting issues because of the autonomy feature of agents and also because of the environment changes. A model for multi-agent planning should be studied from several perspectives: dynamic (or reactive) planning, decentralized planning, task allocation and resource sharing, etc. Decentralized planning is addressing issues for autonomous agents (e.g., autonomous vehicles) cooperating in complex missions (e.g., in space) that often involve a set of tasks, each representing a component of the mission (Nam & Rhee, 2010). Task planning algorithms are used as part of the mission planner to assign agents to tasks.

2.10 Conflict resolution

In general, conflict resolution can be considered as the set of methods and processes involved in discovering and applying solutions for eliminating conflict between any kinds of agents. Committed, within a specific workgroup, agents attempt to resolve workgroup conflicts by actively communicating information about their conflicting motives or ideologies to the rest of the group (e.g., intentions; reasons for holding certain beliefs) and by engaging in collective negotiation.

As already stated, autonomous agents plan their paths through known and unknown environments to reach their goals. When multiple autonomous agents share the same area, conflict situations may occur that need to be solved. There are several approaches towards solving conflict resolution in autonomous agents (e.g., (Alshabi, Ramaswamy, Itmi, & Abdulrab, 2007; Düring & Pascheka, 2014; Jacak & Pröll, 2009)), with an emphasis in the decentralized decision making algorithms (Düring & Pascheka, 2014) based on two main ideas: Gathering individually available manoeuvres from other agents and choosing a cooperative manoeuvre combination. Such an approach combines the use of motion primitives to define a set of available manoeuvres with an intelligent decentralized decision making algorithm that is based on the operationalization of cooperative behaviour (ability of working together towards increasing the overall performance of the workgroup).

Applications of intelligent controllers solving conflict situations by coordinating autonomous agents include robotics, unmanned aerial and marine vehicles, air traffic management, automotive engineering. A conflict resolution mechanism is needed to let the vehicles determine the crossing order and optimal trajectories by themselves. A distributed conflict resolution mechanism that does not depend on a traffic manager is highly desired for small intersections and in the case that the manager is broken. A communication-enabled distributed conflict resolution mechanism is required in order for a group of connected autonomous vehicles to move safely and efficiently in intersections in the absence of a traffic manager (Liu, Lin, Shiraishi, & Tomizuka, 2018). Decision making (for the determination of passing order) as well as motion planning (for the computation of trajectories), are both required.

2.11 Reasoning

Reasoning is related to humans’ thinking, cognition, and wisdom. In AI, reasoning is divided into logical reasoning types such as deductive reasoning, inductive reasoning, abductive reasoning, and other less formal methods such as intuitive reasoning and verbal reasoning. Logical reasoning is about how agents understand information gathered by sensors within their environments, or how they conceptualize cause, effect, truth, or even how they create ideas regarding the notions of good and bad. Automated reasoning

is considered a subfield of AI with connections to theoretical computer science and philosophy. Formal logic played a major role in the field of automated reasoning, which itself led to the development of AI.

There are different kinds of AI reasoning in autonomous agents e.g., case-based reasoning (Vacek, Gindele, Zollner, & Dillmann, 2007), probabilistic reasoning (Lance Eliot, 2018). Vehicle guidance in complex scenarios (e.g. inner-city traffic) requires an in-depth understanding of the current situation of the vehicles and environment they interact (other vehicles, pedestrians, obstacles, etc.). To select the appropriate behaviour for an autonomous vehicle, a deep and real-time analysis of the current situation is needed. Such an analysis may consist of an estimation of the situation's evolution with respect to the selected behaviour. Case-based reasoning allows to utilize prior experiences for the assessment of situations. As in human driving, where there is a continuous adjustment of probabilities due to the obvious uncertainty of the environment, one would expect that self-driving cars would perform similarly. Today, most of the self-driving cars that are developed do not integrate probabilities into their AI systems, since it is not as easy as it might seem at a first glance (Lance Eliot, 2018).

2.12 Big Data

Data is key to autonomous agents' technology (Daniel Matthews, 2018). For instance, for an autonomous vehicle to plan and act, it is using machine learning algorithms which need high volume of data to be able to predict outcomes accurately (e.g., a pedestrian is getting ready to cross the street, and the vehicle has to be trained well enough to be able to react appropriately). Another example is real-world data from vehicles out on the road. Such vast amounts of data are not only data from sensors and from the available infrastructure, but also data from crowd-sources as well as personal data from drivers and passengers (Lynnette Reese, 2017). Crowd-sourced data include the reporting of new obstacles or construction. Connected self-driving car systems may upload their data to the cloud and share their data that will be then used to train the systems of other vehicles.

3. Security and privacy considerations

Autonomous agents are characterized by diversity, with applications varying from digital assistants residing in our smartphones to autonomous robots supporting supply chain. Depending on the level of autonomy and the context of operations security and privacy considerations may vary. However, an overview of most common considerations identified is briefly presented below, based on a set of indispensable baseline security requirements as published by ENISA in 2016[8].

3.1 Security considerations

Detection of rogue or unauthorized autonomous systems

Malicious autonomous agents can masquerade as legitimate agents to avoid detection. Developers shall be able to provide means to ensure that the agent has indeed been authorised to perform its tasks, the confidentiality, integrity and availability of data processed is preserved, depending on the operation and the context, the agent cannot be tainted during its operation and its integrity is preserved and verifiable throughout its lifecycle.

Hijacking and misuse

Autonomous drones, vehicles, robots and other devices are controlled by software systems. The degree of vulnerability varies depending on the approach and quality of the software development process and the breadth and depth of testing. The developer shall be able to provide evidence that a managed security by design approach has been adopted, including documented secure software development, quality management and information security management processes.

Interference

Autonomous systems rely heavily on their sensory ability to perceive their environment, which incurs a security risk from sensors. Security researchers (Yan et al., 2016) have examined the security of sensors used to guide self-driving cars to find that they are vulnerable to contactless attacks. The provider shall design and pre-configure the delivered product such that functionalities are based on well-established security practices and are reduced to the strict minimum required for system operations.

Transparency and accountability

The behaviour of autonomous agents is not fully prescribed in their software code. Their behaviour is uncertain and results as a combination of their software, their training, and their perception of the environment. In most advanced autonomous agents, there is no distinction between the training and operation phases, as training continuous autonomously throughout the lifecycle of the agent. Thus, training is not fully controlled by the manufacturer. The manufacturer however shall be able to offer comprehensive and understandable documentation about the overall design of the agent, describing its architecture, functionalities and protocols, their realisation in hardware or software components, the interfaces and interactions of components with each other and with internal and external services, in order to be able to implement and deploy the agent in the most secure way possible.

Adherence to security principles

Similar to the basic security principles that are for information systems inception, development and deployment of an information system, a set of basic security principles has to be identified, defined and

[8] https://www.enisa.europa.eu/publications/indispensable-baseline-security-requirements-for-the-procurement-of-secure-ict-products-and-services

adhered to through the lifecycle of an autonomous agent. Therefore, Components of the agent, services that it uses and services that it provides shall be secure:

- **By design** – the agent, or service, has been conceived, designed and implemented to ensure the key security properties are maintained: availability, confidentiality, integrity and accountability;
- **By default** – the agent, or service, is supplied with the confirmed capability to support these security properties at initial deployment;
- **Throughout its lifecycle** – security should be maintained from initial deployment through maintenance to decommissioning;
- Each of the above principles should be **verifiable**.

3.2 Privacy considerations

Pervasiveness and the minimization principle

The learning capabilities of autonomous agents are based on machine learning algorithms that by nature require large datasets as input. This means that they tend to collect as much data as possible, rather than a statistically relevant sample (Butterworth, 2018). Consider, for example, the case of autonomous vehicles. They are equipped with sensors that collect information from an area with radius of approximately 300 meters. They are equipped with cameras and microphones, as well as radars and laser sensors that create high-resolution 3D representations of objects in the environment. Autonomous agents do not seem to currently have the capacity to distinguish necessary from unnecessary data. The ability to act autonomously depends strongly on the ability to perceive the environment. As the environment is complex and dynamic, it is difficult to know a priori whether a piece of information will be useful or not.

Data retention and protection

Autonomous agents process real-time data that could be deleted after use without affecting their performance. However, these data are stored and retained for large periods of time for different reasons. One such reason is the investigation of accidents and other incidents. For example, in the case of autonomous vehicles it is very important to be able to investigate an accident, in order to identify the cause and provide a solution/patch. Training data sets are also retained for the same reason as it should be possible to investigate whether they were sufficient and appropriate. Data are also stored and used for research purposes, as they constitute a valuable input in the process of advancing the autonomous agents technologies.

Even when data are deleted, they leave traces that can be exploited. Researchers (Shokri et al., 2017) have shown that it is possible to determine whether a specific record was part of the training set of a machine learning algorithm, without having any access to the internal structure of the model (i.e., black box analysis). They have thus shown that machine learning models are vulnerable to inference attacks. Therefore, we should not consider the deletion of the training set to be an effective measure for privacy protection. The above data sets are accessible by engineers, developers, researchers, investigators and, possibly, many other stakeholders. As a result, the protection of data confidentiality is quite a challenging task.

Data aggregation and repurposing

In most of the applications of autonomous agent technologies data are aggregated by providers and used for several purposes. Autonomous vehicles and drones are connected to control centers owned by the companies that produce these vehicles and drones and data are transmitted for analysis to these centers. In a similar manner, software agents, such as intelligent assistants, share data with their manufacturers that analyse and process them to provide additional services to their users. Developers and providers collect data from numerous users and combine data from various sources. They can create detailed

profiles of people, they can make predictions about their behavior and guess their needs through data aggregation and profiling.

Data collected from one activity can be used and analysed for a wide range of purposes. This is often referred to as “repurposing”. Data that were primarily collected to be used for agent training could also be used for marketing purposes, data collected for the purpose of investigating accidents, could be used for user profiling, and so forth.

The opacity of processing

Traditional data processing systems comprise software that implements well-defined algorithms, such as decision trees. Data are stored separately in rigid structures, such as relational databases. Data subjects and data protection authorities can query databases and inspect processes and procedures to identify how a processing operation operates. Thus, the lawfulness, fairness and transparency of personal data processing can be observed.

On the contrary, machine learning processes operate as a “black box” to the user (Butterworth, 2018). Machine learning algorithms provide no explanation for their results. They cannot show whether a certain data instance has been necessary to achieve the purpose of processing or not and they cannot show the degree of significance of each piece of information. Thus, in the case of autonomous agents that base their autonomy on machine learning modules, it is currently challenging, to demonstrate the lawfulness, fairness and transparency of personal data processing. Hence, there is need to apportion responsibilities across different regulatory and enforcement agencies involved and determine competences across them.

4. Shaping a framework for policy development in cybersecurity

As a first step towards sketching a point of reference for policy-makers tackling the issue of cybersecurity and establishing a relevant framework for policy development, this study provides an overview of representative application domains of autonomous agents and attempts to complement relevant initiatives at EU level by providing insights and considerations, both for security and privacy. The rapid adoption of digital technologies has enabled many opportunities for innovation and interaction but at the same time introduced a number of challenges that were not evident before. The cyber domain has evolved so swiftly that legal, economic, and societal mechanisms for maintaining cybersecurity have struggled to keep up.

Artificial intelligence (AI) and increasingly complex algorithms currently offer great potential and possibilities for a diverse range of application areas which influence our everyday lives more than ever before. Similarly, Industry 4.0 with the Internet of Things (IoT) at its core has already exerted an impact on society, transforming products, customer experience and the labour market. However, is not viable to expect that each technological advancement with a dedicated legislative instrument, mainly due to the fast pace of innovation. Therefore, a more broad approach must be pursued, able to balance the priorities of all involved stakeholders while putting forward participial solutions to emerging challenges.

In 2016, RAND published a study on "A Framework for Exploring Cybersecurity Policy Options"[9] which aimed to "to develop an initial framework for cybersecurity that considers the roles of government, industry, advocacy organizations, and academic institutions and how these stakeholders' concerns relate to each other. The main findings of the study pertained the lack of cybersecurity demand in the market, the limited incentives to promote cybersecurity best practices and finally the need for public-private partnerships able to identify, elaborate and eventually address emerging cybersecurity challenges.

Based on the analysis of numerous application domains of artificial intelligence in Section 2, it is evident that a framework for policy development in cybersecurity should adhere to two main principles, namely:

- **Inclusiveness**: reflecting and promoting interests and priorities from a wide range of stakeholders;
- **Openness**: supporting the inclusion of future innovative digital technologies in diverse application domains.

Towards a more practical interpretation of these two principles, such a framework should:

- Promote the development and adoption of technical standards and specifications;
- Encourage exchange of best practices and experience;
- Facilitate the establishment of synergies and partnerships between private and public sector stakeholders.

[9] https://www.rand.org/content/dam/rand/pubs/research_reports/RR1700/RR1700/RAND_RR1700.pdf

5. Conclusions and Recommendations

One of the key aspects in autonomous systems is the data collected, mainly for supporting the demanding functionality in a qualitative and timely manner. Given its distinctive characteristics, an envisioned policy framework will require forward thinking practices, both by private and public sector, and ensure that effective approaches are promoted while appropriate safeguards, requirements and obligations are also in place. Towards this direction, the main challenges identified are briefly presented below, in addition to respective recommendations, which are meant to complement ongoing work and initiatives at EU level.

5.1 Adoption of security and privacy by design principles

One of the key aspects in autonomous systems is the ability to process vast amounts of data and act upon them, mainly for supporting the demanding functionality in a qualitative and timely manner. However, the autonomous system must be conceived, designed and implemented in a way that ensures adherence key security properties availability, confidentiality, integrity and accountability. In addition, it must be supplied with the confirmed capability to support these security properties at installation throughout it's lifecycle – from initial deployment through maintenance to decommissioning.

European Commission, relevant EU Bodies and relevant public and private sector stakeholders should further promote and support the adoption of security and privacy by design principles as a pre-requisite during the inception, design and implementation of autonomous agents and systems.

5.2 Development of baseline security requirements

The rapid developments in the fields of artificial intelligence and autonomous agents reinforces the need for a coordinated approach and guidance during their inception, design and development. As such, developers could benefit from guidance on how to address identify and overcome identified security and privacy challenges. Such guidance could be provided through identification and exchange of best practises, development of baseline security requirements voluntary standards which will also provide guidance on conformity assessment.

European Commission, relevant EU Bodies, public and private sector stakeholders should foster a collaborative approach on identification and exchange of best practices. Gradually such initiatives should put forward sets of baseline security requirements which can then be transposed to widely accepted technical specifications and standards.

5.3 Coordination of actions on highlighting and addressing ethical considerations

Autonomous agents and systems often need to make complex decisions, which apart from the aspect of accountability, introduce also ethical considerations with regard to human rights, human dignity and non-discrimination. Addressing ethical issues and dilemmas in autonomous systems design requires an interdisciplinary and coordinated approach.

The cost of coping with machine autonomy is also an issue of concern as it far outpaces the ability of a single community to deal with the entire range of issues alone. The Internet has already challenged societies that have embraced it as effort and resources are dispensed off locally to counter threats originating from afar. In cases such as applications of machine autonomy it is likely that such issues will be

exacerbated further leading to a new set of policy challenges that might exhaust the capability of existing agencies to respond.

During the 40th International Conference of Data Protection and Privacy Commissioners, a declaration on Ethics and Data Protection in Artificial Intelligence was signed[10]. Among others, the declaration endorses a set of guiding principles in the development of artificial intelligence, and as an extension to autonomous agents, while establishing a working group on Ethics and Data Protection in Artificial Intelligence. Similar initiatives should be further encouraged and promoted, both at EU and International level, to ensure that future developments in the area take into account promotion and protection of human rights.

European Commission, relevant EU Bodies, public and private sector stakeholders should further promote and support existing initiatives on promotion and protection of human rights through establishing appropriate ethical parameters.

[10] https://edps.europa.eu/sites/edp/files/publication/icdppc-40th_ai-declaration_adopted_en_0.pdf

6. Bibliography/References

Alshabi, W., Ramaswamy, S., Itmi, M., & Abdulrab, H. Coordination, cooperation and conflict resolution in multi-agent systems. Innovations and Advanced Techniques in Computer and Information Sciences and Engineering, 2007, pp. 495–500.

Arar, S. Facial Recognition Increases Autonomous Vehicles Safety - News. 2016.

Butterworth, M. The ICO and Artificial Intelligence. The Role of Fairness in the GDPR Framework. Computer Law & Security Review, 34, 2018, pp. 257-268.

Daniel Matthews. Data Is Key to Autonomous Vehicle Technology, And Tesla Says It's Winning, 2018.

Düring, M., & Pascheka, P., *'Cooperative decentralized decision making for conflict resolution among autonomous agents'*. International Symposium on Innovations in Intelligent Systems and Applications INISTA 2014, IEEE, 2014, pp. 154–161.

Fallah-Seghrouchni, A., Multi-agent Planning for Autonomous Agents' Coordination. In Monitoring, Security, and Rescue Techniques in Multiagent Systems, Berlin/Heidelberg: Springer-Verlag, 2005, pp. 53–68.

Fox, J., & Khan, O., From decision support systems to autonomous agents: how can we ensure ethical practice? Doc.Gold.Ac.Uk, (April), 2014.

Garbhe, S. What is Artificial Intelligence (AI) – Becoming Human: Artificial Intelligence Magazine, 2018.

Hess, T. J., Rees, L. P., Matheson, L. A., & Ragsdale, C. T., Study of Autonomous Agents in Decision Support Systems, 1999.

Hess, T. J., Rees, L. P., & Rakes, T. R., *'Using Autonomous Software Agents in Decision Support Systems. In Handbook on Decision Support Systems 1'*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 529–555.

Hirz, M., & Walzel, B., *'Sensor and object recognition technologies for self-driving cars. Computer-Aided Design and Applications'*, Vol. 15 No. 4, 2018, pp. 1–8.

Jacak, W., & Pröll, K., *'Conflict Resolution in Multiagent Systems Based on Wireless Sensor Networks'*. Springer, Berlin, Heidelberg, 2009, pp. 753–760.

Lance Eliot. Probabilistic Reasoning for AI Self-Driving Cars - AI Trends, 2018.

Liu, C., Lin, C.-W., Shiraishi, S., & Tomizuka, M., *'Distributed Conflict Resolution for Connected Autonomous Vehicles'*. Transactions on Intelligent Vehicles, IEEE, Vol. 3 No.1, 2018, pp. 18–29.

Lynnette Reese. Big data in autonomous driving | Solid State Technology, 2017.

Nam, M., & Rhee, P. K., *'Agent Based Object Recognition System for Autonomous Surveillance in Dynamic Environments'.* 3rd International Conference on Human-Centric Computing, IEEE, 2010, pp. 1–5.

Nami, M. R., & Bertels, K., *'A Survey of Autonomic Computing Systems'*. International Conference on Autonomic and Autonomous Systems ICAS'07 IEEE, 2007, pp. 26–26.

Poslad, S., Autonomous Systems and Artificial Life. In Ubiquitous Computing. Chichester, UK: John Wiley & Sons, Ltd., pp. 317–341.

Russell, S. J., Stuart J., Norvig, P., & Davis, E., Artificial intelligence: a modern approach. Prentice Hall, 2010.

Salomon, R. (1996). *'Neural Networks in the Context of Autonomous Agents: Important Concepts Revisited'*. Artificial Neural Networks in Engineering ANNIE'96, 1996, pp. 109–116.

Shokri, R., Stronati, M., Song, C. and Shmatikov, V., 2017, May. Membership inference attacks against machine learning models. In Security and Privacy (SP), 2017 IEEE Symposium on (pp. 3-18). IEEE.

Vacek, S., Gindele, T., Zollner, J. M., & Dillmann, R. '*Using case-based reasoning for autonomous vehicle guidance'*. International Conference on Intelligent Robots and Systems, IEEE/RSJ, IEEE, 2007, pp. 4271–4276.

Yan, C., Xu, W., & Liu, J. (2016). Can you trust autonomous vehicles: Contactless attacks against sensors of self-driving vehicle. DEF CON, 24.

ENISA

European Union Agency for Network
and Information Security
1 Vasilissis Sofias
Marousi 151 24, Attiki, Greece

Heraklion Office

Science and Technology Park of Crete (ITE)
Vassilika Vouton, 700 13, Heraklion, Greece

Catalogue Number TP-06-18-399-EN-N

1 Vasilissis Sofias Str, Maroussi 151 24, Attiki, Greece
Tel: +30 28 14 40 9710
info@enisa.europa.eu
www.enisa.europa.eu

ISBN: 978-92-9204-282-0
DOI: 10.2824/453043