Artificial Intelligence, Security and Rights

Filipe Fernandes

Instituto Superior Técnico, Lisbon, Portugal

Abstract

Artificial Intelligence (AI) is a technology with the potential to transform our life. It appears to be the answer to several societal challenges that we face, and it is poised to be a real engine of economic development. However, we must properly study and address the legal, ethical, and socio-economic concerns that arise from its development and use.

The present study aims to identify the effects of AI systems on citizens' security, data, and freedoms, as an integral part of society.

Firstly, we give an overview of the concepts needed to understand the context and the inner workings of AI, then we studied concrete examples of AI applications in the security domain – specifically in the domains of Cybersecurity, Predictive Policing and Video Surveillance using Facial Recognition.

There are clear indicators that AI will bring added efficacy and efficiency to the processes that the systems address, but we also concluded that there may be interference on citizens' freedoms, rights and guarantees in the use of AI systems. We reflected on the measures that assure the development and use of AI systems in a responsible, ethical, and safe manner.

The challenge, presented to us, requires us to follow a path of awareness of the capabilities and impacts of the development of AI, identifying guidelines to ensure respect for the citizens' freedoms, rights and guarantees, while benefiting from the fruits of this technology.

Keywords: Artificial Intelligence, Security, Rights, Cybersecurity, Predictive Policing, Facial Recognition.

1 Introduction

1.1 Investigation problem

The objective of this research is to study the emergence of Artificial Intelligence (AI) as a security vector and the implications of its use in a society that aims to be democratic and in full respect of its citizen's fundamental rights, such as privacy, freedom of expression and freedom from discrimination.

The challenges that our society faces regarding the rise of AI should be subject of a broad societal discussion, separating facts from fiction and increasing awareness of the capabilities and repercussions of this new and emerging technology.

The scientific community continues to give warnings about the potential dangers that this technology entails. Especially when it develops, matures, and reaches a level of intelligence equivalent to that of the Human being, or even higher, something that seems distant but is real (1).

We consider the rise of General AI a major challenge for humanity and a subject that should widely discussed and studied. However, in this work we focused on narrow applications of AI, designed to solve specific problems, and the immediate consequences already observable.

Despite many technological innovations, security will remain a fundamental need for society. The security challenges that marked the beginning of this century, such as terrorism and cyber threats, have also undergone evolutionary processes with the advancement of technology. This evolution forces the nations and their Law Enforcement Agencies (LEA) to study these phenomena, adapt and give an adequate and effective response.

This study will address the main question:

MQ: How will the use of AI systems affect our security, our freedom and privacy?

We study the main concerns that arise from the use of this technology on citizens as part of a society.

We also defined the following derived questions:

DQ1 - What AI systems can support LEA?

DQ2 - What interferences can occur in citizens' rights and freedoms, with the use of AI systems by LEA?

DQ3 - What measures should be taken to ensure the development and use of AI systems by LEA is done in a way that is responsible, ethical, and safe?

Answering the questions obliges us to follow a path of awareness of the capacities and impacts of the development of AI, identifying guidelines to ensure respect for the rights, freedoms and guarantees (RFG) of citizens.

2 Core Concepts

2.1 Security

Security is a concept of a polysemic nature and in constant evolution. We consider that a comprehensive definition encompasses the concepts of public security, public safety, crime prevention and investigation, which in turn are absorbed by a broader concept, that of national security (2).

Also important is the notion of "right to security", which materializes the guarantee of the exercise of any rights freed from aggression or threats. Security means two things: "the right of defence against aggression by public authorities and the right of protection conferred by public authorities against aggression or threats by others." The main visible driving force of security are LEA and to understand this concept we must develop the concept of police. Police can be defined as a way of acting by the administrative authority, which consists in stopping the exercise of individual activities that may endanger general interests and prevent the social damage that the law seeks to protect." (3).

2.2 Artificial Intelligence

AI is reconfiguring our society starting with the way we relate, the way we work and even our economy, promising to generate productivity gains, reduce costs and make all processes more efficient.

AI is a branch of Computer Science whose purpose is the study of intelligent agents: any device that recognizes its environment and performs actions that maximize its probability of successfully reaching its goals. (4)

AI systems are comprised of software, and sometimes hardware who, having received a complex objective, act in the physical or digital dimension. They perceive their environment through the acquisition of data, either structured or unstructured. The processing of collected data results in information, which is used for reasoning and understanding the environment. This will be applied in a decision which is best suited to achieve the established objective (5).

One main goal of AI is to teach computers how to do the things that humans currently do best, and learning is undoubtedly the most important of these things: without learning, no computer can keep up with a human for a long time (1).

In recent years, significant advances have come from the sub-discipline of AI, Machine Learning (ML) and even more from a sub specialization called Deep Learning (DL), which focuses on teaching machines by applying algorithms to data. The terms AI, ML and DL are often used interchangeably in an erroneous way.

This process called ML takes many different forms and is known by many different names: pattern recognition, statistical modelling, data exploration, knowledge discovery, predictive analysis, data science, adaptive systems, self-organized systems, and much more. Each of these is used by different communities and has different associations (6).

All algorithms have an input and output: the data enters the computer; the algorithm does what it wants with them, and the result comes out. ML reverses this situation: we insert both the data and the desired result; the algorithm learns how to transform the first into the second (6).

This approach requires vast amounts of data from a specific domain to be able to optimize the decision-making process to achieve the desired objective. This is can be used to recognize deep patterns and correlations that link many information points to an objective (1).

3 Field Research

In this chapter, we address the first Derived Question: *What AI systems can support LEA?*

There are general applications of AI to areas such as human resources management, internal process management, management of work environment. We focused our interest in applications exclusive to the security domain. There are several AI applications and systems under development and in use, such as: voice to text translation, analysis of telecommunications, text analysis for information production, virtual agents to collect testimonials or receiving complaints, autonomous land and air vehicles for border patrol, identification of tax fraud, and identification of publications with prohibited content on social networks (7).

LEAs carry out several missions, each corresponding to a potential application of AI. We analysed three areas of special interest with promising developments: cybersecurity, predictive policing (PP) and video surveillance using facial recognition (FR).

3.1 Cybersecurity

The preservation of confidentiality, integrity and availability is the holy grail of information security in cyberspace. In turn, cyberspace is the complex environment resulting from the interaction of people, software, and services on the Internet through devices and networks.

The area of cybersecurity is a notable example of AI applications. Particularly in defensive applications such as: Malware classifiers, with similar functions to anti-virus (AV); threat intelligence; behavioural analysis to detect insider threats in organizations; and the creation of adaptive Honeypots that simulate a digital "shadow organization" to lure attacks.

The traditional AV approach to detect malware is by comparing a digital signature, extracted from the suspect file, with the signatures of known malware samples in the AV database. Several companies have started to develop AI software to detect, investigate, classify, and mitigate the most advanced and unknown types of malware in a preventive and real-time manner.

We studied HP Sure Sense, which was developed by HP with DeepInstinc. This software uses a previously trained detection model, with data from hundreds of millions of files, classified as safe or malicious. During the training process, the algorithms defined the characteristics or attributes that differentiate a safe file from a malicious file to create an AI prediction model that works like a classifier.

Each file that tries to interact with the device or is accessed by it, is verified by the agent, and receives a score. The score represents the level of maliciousness of the file. If that value is above the threshold established for a safe file, the agent prevents its execution. In addition to classifying whether it is safe or not, it also classifies in real time the type of threat or family of Malware.

We also studied IBM X-Force Exchange, which is a tool for sharing threat intelligence – information about cybersecurity threats. This can be defined as the collection of evidence about a cyber-attack, be it its context, mechanisms of the attack, recommended actions, or other information necessary to support the Security Operations Centre (SOC) personnel.

The "intelligent" part of this tool is the ability to leverage its natural language processing (NLP) capabilities to digest vast amounts of structured and unstructured data, such us malware signatures, IP addresses, security reports, websites and so on, and produce relevant information to the investigation of cybersecurity incidents.

It generates information reports that give a summary of the vulnerability or threat, a detailed description, indicators of compromise, digital signatures of the Malware and recommendations on how to mitigate and respond to such threats.

These tools contribute to the security of organizations, by increasing the protection of information systems, detecting malware more effectively and by facilitating the dissemination of knowledge among elements that deal with cyberspace threats daily.

3.2 Predictive Policing

Several policing models represent the different approaches that States, and police adopt to tackle crime. Some are more reactive, others more proactive, but all aim to obtain legitimacy and success in their actions and to win the citizens' trust.

PP is the application of analytical techniques – mainly quantitative techniques – to prevent the occurrence of crimes, to identify potential victims or criminals through statistical predictions (8).

According to Ratcliffe: "Predictive policing, is the use of historical data to create a spatial-temporal forecast of potential crime areas and critical points. These will serve as a basis for decision-making regarding the allocation of resources, in the expectation that the police force will be at the time and place of the criminal occurrences." (9).

In recent years, it has been possible to observe an increase in interest, demand and investment in analytical tools that use large data sets and big data to make predictions in support of criminal prevention.

The main objective is to increase the effectiveness and efficiency of LEA, seeking to move from a reactive posture to a proactive attitude, leveraging its limited resources.

"By predicting crime trends and by strategically concentrating patrols where they are most likely to be needed, as well as taking other preventive measures, the police force in one city can actually do the job of a much larger one." (6).

We studied PredPol and Hunchlab, two patrol management platforms used for command and control of LEA assets, mainly patrols.

These platforms develop models that consider different traditional prediction tools such as: criminal indices and hot spot maps; contagion, relative to the spread of recent events; modulation of the terrain-risk relationship, proximity, and geographical density of points of interest; routine activity theory, which integrates the actions of criminals, LEAs and potential victims; collective effectiveness, using socioeconomic indicators, time cycles, related to seasonality, time of day, week and month, recurring events, holidays, sports seasons, and finally weather conditions, such as temperature, precipitation, etc.

The software presents these predictions designating areas where crimes are most likely to occur. Patrols are advised to spend at least 15 minutes in each square engaging in different tactics depending on the predicted crime. Making predictions is only a part of the solution. It is up to the LEAs to adopt the best police tactic for the given situation or area referenced by the tool.

3.3 Video surveillance using facial recognition

This AI application shows incredible potential in finding missing children, identifying, and tracking criminals and alerting when suspicions objects are abandoned. On the wrong hands, it may constitute the first step towards realizing an Orwellian prophecy.

FR is a form of biometric recognition, as the face has several distinctive features that can be measured and translated into a unique identifier, which can be efficiently verified. This unique identifier can be converted into digital format to allow its storage and search (10).

FR can be used for identifying someone by comparing its face with a database of known people's faces, designated, one-to-many. It is also used for identity verification, designated one-to-one. The one-to-many recognition process generally consists of two phases: first the pre-processing phase and secondly the matching phase.

During pre-processing, the photo of a known person is scaled, aligned, and processed by the FR Software with several possible scanning methods. Facial features are quantified and mapped over different masks representative of the facial structure and can be stored in different formats as a representation of an individual's face, a facial impression (11).

This facial impression is stored with biographical information of the individual in the database of known people. During the correspondence phase, the FR Software processes a photograph or frame of the person to be identified, obtaining its facial impression. This is compared to the facial impressions present in the database of known persons.

The facial recognition software uses an algorithm to compare facial impressions that have a similarity value between various possible matches.

If the Software determines that match surpasses the similarity threshold, they will be presented as a probable match. Depending on the software, it identifies one or more likely matches that should be validated by a human being (12).

In this study, we analysed a live FR matching software from Hitachi Vantara, which is capable to match up to 60 faces per second from a single server with four video feeds. Additionally, it could set up virtual fence alert systems, analyse traffic, count vehicles, and read their license plates, detect intrusions, detect objects such as weapons or barriers, and can be used for parking lot management. This product is a true force multiplier, having virtual agents watching over videos feeds, which would otherwise be impossible to monitor and tedious to analyse in search of suspects. It shortens the time identifying and locating threats, it helps criminal investigations tracking criminals, making police work more effective.

4 Discussion

4.1 AI Interferences

After studying different AI system’s capabilities and practical applications, we consider the second Derived Question: *What interferences can occur in citizens’ rights and freedoms, with the use of AI systems by LEA?*

There are already several AI systems in use that raise ethical concerns, such as, citizen score system, communication monitoring systems, autonomous lethal weapon systems and long-term concerns with General IA development (13).

The use of AI for cybersecurity by LEAs has reduced the impacts on citizens' rights, as they are defensive tools. On the other hand, it can lead to a loss of privacy because of the added capacity to analyse large amount of data. Moreover, it is foreseeable that AI will be used, by state or non-state actors, for offensive or malicious purposes.

There are two distinctive categories of malicious use of AI: AI-supported attacks, those that include AI-based techniques aimed at improving the efficacy of traditional attacks, and AI-targeted attacks, focused on subverting existing AI systems to alter their capabilities. Some probable use cases are the creation of more sophisticated types of attacks, e.g., AI powered malware with the capacity to morph and adapt to the environment where it is deployed. Advanced social engineering, with massive spam attacks with highly tailored victim related knowledge and information. AI-powered fake social media accounts farming, with human interaction simulated by AI systems. Deep generative models to create fake data for poisoning AI training model's datasets (14).

The lack of regulation of the use of AI can lead to a proliferation of cyber weapons. It is urgent to define limits to its use for defensive and offensive purposes. Identify what are the legitimate or illegitimate targets. Demand rigorous assessments of proportionality in its use by States and International Organizations (UN, NATO, EU). Define rules and principles that legitimize the intervention of these bodies under national and international law.

The second area of discussion is Predictive policing programs. They can be directed to two types of crimes, violent crimes – which include homicide, arson, theft, and assault, which are usually reported. The second type is petty criminality – which includes the sale and consumption of small amounts of narcotics, driving without license, damages, theft, or perjury (15).

Many of these crimes, sometimes referred to as antisocial behaviour, would not be registered or reported if they were not witnessed by targeted patrolling. This petty crime is endemic to many poor neighbourhoods.

Cathy O'Neil finds it unfortunate that the tools are being directed at this type of crime. "Including them in the model threatens to skew the analysis." Once the nuisance data flows into a predictive model, more police are drawn into those neighbourhoods because they are evaluated as more prone to the occurrence of crimes. These patrols can initiate a feedback loop, feeding the model with even more information from these areas and resulting in a system that promotes excessive policing and marginalization of society (16). This program may affect disproportionately the most vulnerable population being counterproductive.

Facial recognition is a diffuse technology that lacks regulation in the algorithms and oversight of its use. This, in turn, can lead to racial prejudices and serious social consequences. Facial recognition is a more efficient and less invasive form of biometric identification. However, one study proved a case of algorithmic discrimination related to facial recognition. This study evaluated the effectiveness of three commercial facial recognition products including Microsoft, IBM, and Face ++. The study revealed the following conclusions: all classifiers perform better on male than female faces (difference of 8.1% to 20.6% in error rate); all classifiers perform better on lighter faces than darker faces (11.8% to 19.2% difference in error rate); all classifiers perform worse on darker female faces (20.8% to 34.7% difference in error rate); the Microsoft and IBM classifiers perform better on lighter male faces (error rates of 0.0% and 0.3%, respectively); the Face ++ classifier perform better in darker male faces (error rate 0.7%), the maximum difference in the error rate between the best and worst ranked groups is 34.4%. The study's findings reveal that the algorithms may have different performance rates depending on gender or skin tone, confirming one of the first cases of algorithmic prejudice (17).

This is a consequence of using training data set that is limited and misrepresents reality. The algorithms did not have enough examples to train their skills on the less common face types in the training data sets. Recently, several researchers have published training data sets that are more representative of the population. When used, they help the algorithms to obtain better results in facial recognition on different faces.

There are reports of several innocent people who have been detained due to incorrect identification by a facial recognition program. There are also growing concerns with the indiscriminate use of this technology to control people in demonstrations, which can have serious impact in the exercise of freedom of expression and freedom of assembly and protest. This are clear signs that there should be more control and debate about the use of this technology by LEAs.

4.2 Ethical AI

After seeing AI system's potential ethical problems, we consider the third Derived Question: *What measures should be taken to ensure the development and use of AI systems by LEA is done in a way that is responsible, ethical, and safe?*

Awareness of the evolution and reach of AI systems used by the LEA's triggered a process of national and international discussion. We must ensure that its use is beneficial for each citizen as an individual.

We go over the Unified Framework of the 5 Principles for AI in Society, which are the fundamental principles that should guide the development and use of AI systems: beneficence, damage prevention, autonomy, justice, and explainability.

Then we presented the Ethical Guidelines for a Trustworthy AI, which is a framework that seeks to materialize the principles and concerns of society. It is a checklist to develop an AI system with guarantees of compliance. The development of these systems must comply with the requirements of being legal, ethical, and solid. They must respect the ethical principles of human autonomy, damage prevention, equity and explainability – the basis for a reliable AI.

This framework presents seven requirements that must be considered during the development and use of these systems. These are human action and supervision, technical strength and security, privacy and data governance, transparency, accountability, diversity, non-discrimination, equity, social and environmental well-being. These requirements can be met through technical and non-technical methods throughout the entire life cycle of the system, in a continuous and dynamic way.

In addition to the previous initiatives, codes of conduct, standards and norms that guide the use of these systems for specific application cases must also be adopted. As well as the assessment of conformity with standards and certificates that will be developed in the future. To guarantee all the above factors, regulatory and supervisory bodies must be created. Alternatively, the existing bodies must be adapted to oversee these types of systems.

5 Conclusion

We concluded that AI systems are reaching a stage of maturity and diffusion in several areas related to LEA's with different implications in citizens' rights.

In Cybersecurity, it is being applied to malware detection, adaptive honeypots and the collection and sharing of information about threats. The main ethical issue raised in the development of AI in this area is the continuous "arms race", resulting from its use both defensive and offensive. There is also concern for its malicious use for as AI-supported attacks, AI-based techniques aimed at improving the efficacy of traditional attacks and AI-targeted attacks, focused on subverting existing AI systems to alter their capabilities. To mitigate these issues, it is important to implement security and privacy by design. Foster a culture of data protection and respect for privacy and regulate the use of AI for defensive and offensive purposes, especially as stand-alone countermeasures.

In the area of predictive policing there is a lack of studies that prove its effectiveness and concrete effects in the policing practice. Combined with the lack of transparency and explainability of the suggestions generated by the programs, compel us to rethink their development and employment model. One of the main problems with this application is the importance of the data and models used to make the forecasts. There are concrete risks of the reinforcing the prejudices existing in our society. We must ensure that the data used in the training of predictive models are representative and free from prejudice and the algorithms used must be open and auditable. PP programs must be geared towards the most violent or serious crimes and must be used in a methodical and supervised manner. The implementation of these systems must be phased, favouring controlled pilot projects with evaluation and transparency in all stages. In short, we must make it beneficial and explainable.

The gains in effectiveness and efficiency are more visible in video surveillance applications with facial recognition, providing capabilities to minimize response times to crimes. As well as being an excellent tool for criminal investigation and evidence collection. However, the use of this technology may have unpredictable consequences on society. Including social cooling, the limitation of several fundamental rights, such as the right to privacy, freedom of expression, assembly, and demonstration.

To prevent these adverse effects, we defend the creation of an autonomous law that regulates FR, complementary to the GDPR. This specific legislation must provide and legitimize the use of FR and establish the form, principles, and limits of its use. It must provide a catalogue of crimes, especially those serious and harmful, for which FR is allowed. It is also important to make mandatory the judicial decision or authorisation that legitimizes the process and guarantees a strict control over its use.

To ensure that the use of this technology is beneficial for humanity and for each individual citizen, we must encourage discussion and awareness on the subject. We must establish codes of conduct, regulations and standards that guide the use of these systems for specific cases of application, as well as assessing compliance with standards and certificates that are developed and inspected by specialized agencies.

AI has enormous transformative potential, but like any technology there is a possibility that AI will be employed for nefarious uses. Thus, we must remain vigilant to identify the first signs of interference in the citizens' rights. We ought to develop solutions, checks and balances, to mitigate the risks this technology entails. We must guarantee the safe, conscious, and ethical use of AI to truly benefit from the wave of development driven by it.

References

1. Lee, K. AI Superpowers, China, Silicon Valley and the new world order. Lisbon : Relógio de Água, 2018.
2. Pereira, R. Modelos Preditivos & Segurança Pública. s.l. : Fronteira do Caos, 2018.
3. Caetano, M. Princípios Fundamentais do Direito Administrativo. Rio de Janeiro : s.n., 1977.
4. Poole, D, MACKWORTH, A e GOEBEL, R. Computational Intelligence: A Logical Approach. New York : Oxford University Press, 1998.
5. EU High-Level Expert Group on Artificial Intelligence. A definition of AI: Main capabilities and scientific disciplines. Brussels : European Commission, 2019.
6. Domingos, P. The Master Algorithm. Lisbon : Manuscrito, 2017.
7. INTERPOL & UNICRI. Artificial Intelligence and Robotics for Law Enforcement, 2019.
8. Rand Corporation. Predictive Policing and the role of Crime Forecasting in Law Enforcement Operations, 2013.
9. Ratcliff, J. Intelligence-Led Policing. New York : Rotledg, 2016.
10. Woodward, J. et al. Biometrics: A look at facial recognition. Santa Mónica : Rand, 2003.
11. Ricanek, K. e Boehnen, C. Facial analytics: From big data to law enforcement. IEEE Computer Society : s.n., 2012.
12. Introna, L. e Nissenbaum, H. Facial recognition technology: A survey of policy and implementation issues. Lancaster : The Department of Organisation, Work and Technology, Lancaster University, 2010.
13. Smuha, N; EU Comission. Ethics guidelines for trustworthy AI. Brussels: High-Level Expert Group on AI , 2019.
14. ENISA. AI Cybersecurity Challenges, Threat Landscape for Artificial intelligence. Attiki, Greece : European Union Agency for Network and Information Security, 2020. 978-92-9204-462-6.
15. Malheiros, J, et al. Espaços e Expressões de Conflito e Tensão entre Autóctones, Minorias Migrantes e Não Migrantes. Lisboa : Observatório da Imigração, 2007. 97898980000293.
16. O'Niel, C. Weapons of math destruction: how big data increases inequality and threatens democracy. s.l. : Crown, 2016. 9780553418811.
17. BuolamwinI, J e Gebru, T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Conference on Fairness, Accountability, and Transparency. 2018.
18. Moreira, V, Gomes, C e Neves, A. Compreender os Direitos Humanos. Graz : s.n., 2012.