MOBILIZING INSIGHTS IN DEFENCE AND SECURITY

MINDS

MOBILISATION DES IDÉES NOUVELLES EN MATIÈRE DE DÉFENSE ET DE SÉCURITÉ

Department of Electrical and Computer Engineering
Concordia University

TITLE:

Accountability and Transparency in AI Systems

R. Bahrevar

SUPERVISOR:
Dr. Khashayar Khorasani

Abstract

Artificial intelligence allows us to understand the ever-growing data that is produced worldwide. Since AI systems can produce powerful models that can make sense of large amounts of data, this will stimulate the increase of investments in AI-based smart technologies in various areas. Scientific papers, Industry, and job markets are showing significant interest in the development of AI-based technologies. To make sure the regulation of these systems has not failed to keep sight of these advancements, here we address our perspective and recommendations on barriers in ethics, security, fairness, and transparency for the current AI-based technologies as well as the emerging technologies.

We address definitions and ethical constraints regarding an AI system. What type of features should an AI system have to be considered Ethical? What is data privacy? How can a non-transparent AI system create problems with respect to the different applications? How the development of internet networks will affect AI-based technology? Who is accountable? These are the type of questions that we try to answer in this document.

In the first part of this document, we discuss why the technical aspects should be regarded as a criterion in the development of an AI system. Based on the review of important papers in the literature, this report approaches preventive measures that can be considered in terms of standards, which may help us to regulate the AI systems and hold AI developers accountable.

In the second part, we have discussed potential privacy, security, and fairness concerns of the near future where AI system would be combined with a much more advanced internet capability such as 5G technology, edgecloud ethnology, or a combination of these two. How can we benefit from the cutting-edge technologies that this combination has to offer, as well as respond to the privacy concerns that are much more complicated?

Conceivably, by considering the right approach for the AI design, proactiveness, and push for consideration of ethical constraints we can become closer to an "ethical by design" AI.

1 Introduction

Artificial intelligence is becoming a non-separable part of the technology, and in many aspects facilitates its advance at a high rate. It has been widely used in applications such as autonomous driving systems, recommender systems, face recognition, speech recognition, handwriting recognition, computer vision, automated reasoning, and its true potential may yet be unimaginable. Deep learning is one of the most famous subclasses of AI systems. Its underlying principle of decision making is referred to as the black-box model. Even with the knowledge about the mathematics behind the model, it is difficult to construct an explanatory structure [1]. Lack of explainability is problematic in light of the accountability of the AI systems with regards to their resulting decision for different inputs. Decisions in the area such as determining persons weight based on their social media image [2], predicting crime using information from the twitter[3], recruitment AI [4], Business Intelligence (BI) [5], foreign policy decision making [6]. As a consequence questions about ethics, privacy, security, and fairness of the decision-making process will surface, which [7] refers to them as essential characteristics of an explainable AI.

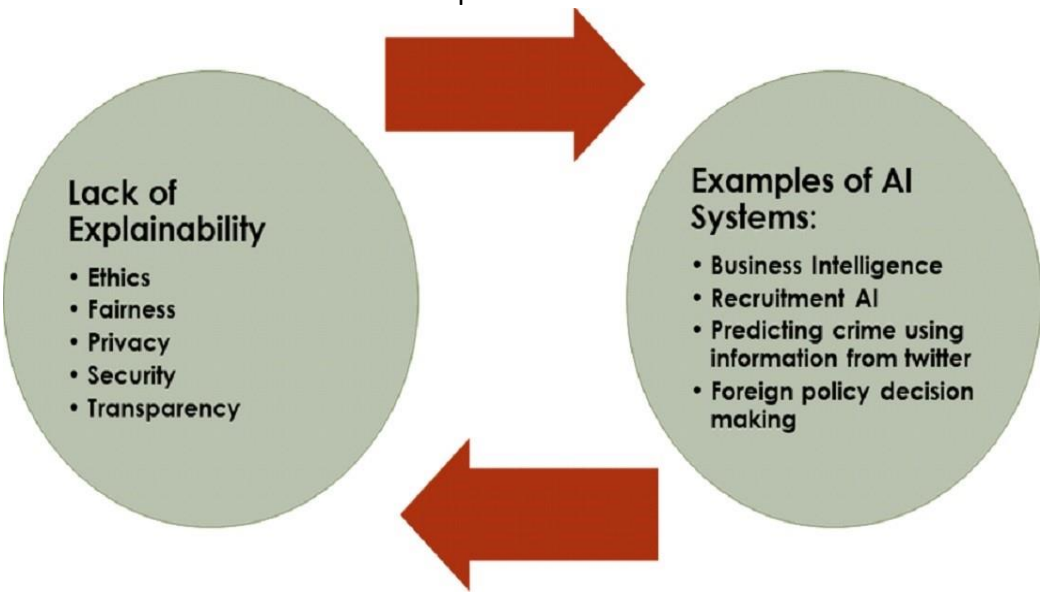

Figure 1: Issues with AI

As mentioned, the characteristics of explainable AI systems are ethics, fairness, security, and privacy, these concepts can be integrated into AI, through introducing new policies and presenting them as standards, or guidelines that would ensure the accountability of AI systems [8]. The ideal situation is to achieve an "Ethical by Design AI" [8]. Here, we argue that there are two viewpoints on the design of an ethical AI, the first one is regarded as the ethical constraint which encompasses all the ethical, policy aspects, and concerns that must be constantly addressed with regards to currents standards on the ethical ground [8], and the other is referred to as the transparency of the methodology. In this report, we try to address achieving the ethical AI through these viewpoints.

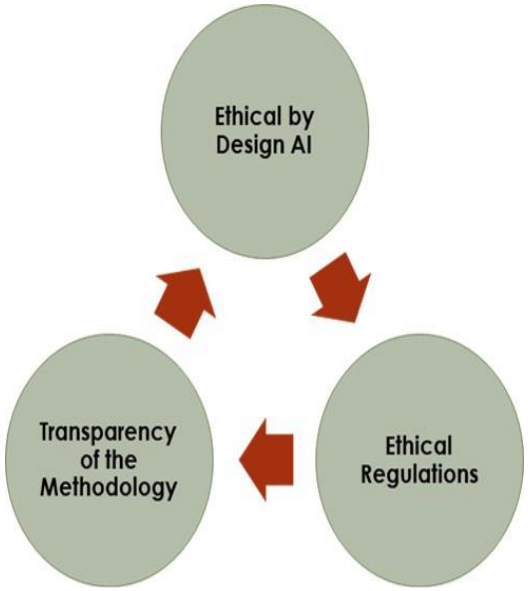

Figure 2: Ethical by Design AI

In terms of ethical/moral grounds, the issues can be viewed in terms of what the general public and we as human beings believe that is morally the right thing, and what steps should we take to prevent undesirable events such as discrimination, loss of private information, defamation, and harm to the general public / the clients dealing with such systems. To deal with this, we have to consider what tools or standards should we look into to be able to come up with the policies, which protects the privacy and serves the interest of the people.

In terms of technical aspects, it is important to readjust our policies based on studies that are concerned with the security and transparency of the AI systems. To address this issue, one approach is to investigate the notion of the explainability for the AI systems.

In [7], a conceptualize perspective to the notion of explainability is presented. Based on this study, AI systems are divided into three categories of opaque systems (e.g. black box models), interpretable systems (e.g. SVM), and comprehensible systems (e.g. ultra-strong machine learning). Opaque systems are defined as the general type of systems where maps between input and outputs are invisible to the user. Interpretable systems are defined as a type of system where users cannot see, study, or understand how inputs are mapped to the outputs. Comprehensible systems mean the systems provide some information, along with the decision that would allow the user to understand why a certain output has resulted from a certain input. Here, they argue that AI systems should be augmented with human-understandable reasoning, such that it encompasses the characteristics of an explainable AI. An interpretable AI system [9] brings a certain degree of trust toward the AI, it is an improvement that will help to facilitate debugging, increase understanding of the AI by the user, decrease subconscious biases, and brings trust toward the performance AI system. Another approach is to investigate methods such as adversarial AI [10] and security certification [11], that deal with the vulnerability of the AI system. As an example, we try to investigate what are the possible threats to an AI system, and what are the possible solutions to counter these threats?

In this report, we try to identify discussions that can be set into policies and can be beneficial to the transparency, security, and accountability of the AI. We also aim to recommend extra measures such as observer design based on explainable AI, which can help with the security and transparency of the AI systems. We want to show how can we utilize these ideas to become closer to an ethical by design AI. The organization of this report is as the following.

In section 2 we discuss a survey of definitions and concepts regarding the characteristics of an ethical by design AI, and in the following, a review of the available techniques for achieving these constraints will be presented. Based on the investigated methodologies and ideas, we present our model for dealing with the accountability issues for the AI system, and express essential element for the technical certification of an AI system and section 6 is the conclusion.

Section 3 is dedicated to policy regarding to new advancement in IoT and how these advancement will effect our privacy and security with regards to AI sensory system.

1.1 Challenges

Artificial intelligence allows us to understand the ever-growing data that is produced worldwide. According to the IDC[1] (International Data Corporation), the Global DataSphere [2] will increase from an estimate of 50 zettabytes in 2020 to 175 zettabytes in 2025 [12]. Since AI systems can produce powerful models that can make sense of large amounts of data, this will stimulate the increase of investments in AI-based smart technologies in various areas. The time is now to be proactive such that we can identify the problems, and provide accountability measures that can regulate these systems.

In this regard, we want to identify what are the values that an ideal AI system should encompass. What features reliable AI systems hold? How can we improve trust regarding the AI systems? and our main question is: how can we regulate AI systems such that they are held accountable not by the trust but through scalable criteria?

In addition, we will witness increased development of the internet of things which is essential due to the ever-increasing amount of transferred data in the world. In this regard, Bell has already launched a 5G network in Montreal, Toronto, Calgary, Edmonton, and Vancouver [13]. Based on Gartner's estimate [14], in 2018, "around 10 % of enterprise generated data is created and processed outside a traditional centralized data center or cloud, the amount that will reach 75 % by 2025". Companies like as IBM (International Business Machine Corporation) are already on the move with the further expansion of the internet network, specifically the combination of 5g of edge computing in IoT, which brings computing and cloud closer to the edge device. IBM describes this combination as "cost-effective, better data control, with faster insights, improved response time, and continuous operation", which will further enhance and expand AI systems and their applications.

[1] IDC: International Data Corporation is a highly recognized organization that helps business executives and investors to make fact-based technology decisions.

[2] The Global DataSphere: It is the summation of all the data, no matter created, captured, or replicated in any given year.

We also want to investigate how emerging development in internet technologies such as 5G and distributed cloud computing will complicate the privacy, security, fairness, accountability, transparency, and ethical issues regarding AI systems, and what policies will help us to be ready for this change?

2 Accountability of AI systems

2.1 Literature Review: Ethical Guidelines and Standardization

Here, we approach different viewpoints regarding ethics in artificial intelligence. In [8], an ethical framework for artificial intelligence in Australia is presented. First, we review the definitions of an unbiased AI, and the tools suggested by this study to establish an initial understanding of an ethical framework's concerns:

2.1.1 Core Principal of An AI System [8]

- Transparency: The public should be informed when an AI system uses information that impacts their life. The developers should ensure that their product is transparent about the type of information it utilizes to make a decision. A trade-off exists, by considering this definition of the transparency, since transparency sometimes can lead to the problem of cyber-security, where people with malicious intent can use the information publicized about the AI to generate the undesirable outcome that can harm the public.

- Generate net benefit: An AI system should be beneficial to the people.

- Do no harm: The negative effect of AI systems that are developed for civilian purposes should be minimized. As an example, COMPASS, an AI sentencing tool is mentioned, which predicts and suggests to a judge, whether or not a prisoner will be granted a parolee. Investigation revealed that this system was highly biased toward African Americans [8, 15].

- Accountability: AI systems designers must be held accountable for the negative effect, lack of security, loss of private information, or harm that may cause to the people. A prime example of where accountability and transparency matters, is the propagation of fake news in today's society,

where an AI platform can facilitate the spreading of fake news, which can have a political and global impact. Therefore, developers should follow strict guidelines so the public can track the source of an undesirable outcome that is influencing their lives. Nevertheless, if an undesirable outcome occurred, then who is responsible? in [8], they argue even the policymaker may be responsible for discrimination caused by AI. Furthermore, it should be the job of a policymaker to consider all the loopholes that an AI developer may use to his advantage and enforce policies that ensure the accountability of an AI system.

- Contestability: When an AI system affects a person's life or community, those affected/ may be affected, by its output, should have the right to argue about its fairness and efficiency. Such platforms should be re-evaluated with the feedback it receives from the public.

  As an example, when an AI system fired teacher in Houston, this matter proceeded to the court, and since there was no transparency on how the software decided to fire the teacher, the court ruled in the teacher's favor [8, 16]. The COMPASS system is also an example of why an AI system must be contestable.

- Fairness: The development or utilization of an AI system shouldn't be discriminative. This is particularly important when the developers are at the stage of training the AI system. Black-box models, especially those directly affecting the public should be trained with data collections, that are fair, unbiased, and represent people from different ethnicity, or community.

- Privacy protection: People's private data must be protected. Here, the prime example is the popular internet platforms such as Google, Facebook, or Instagram, which target us by ads, through analysing and selling information that they collect about our activities on the internet, or they may sell this information for alternative purposes. Privacy means there must be regulation implemented by such companies on the AI systems so that the public can ensure their privacy is not compromised. In a new privacy protection policy set by European commission data regulation, GDPR (general data protection regulation) [17], Europeans can

take control of their privacy setting. Previously, users' consent to the authorization of the personal data for business purposes to an AI system was part of a long agreement, and denying the agreement caused denying the service offered by the AI platform. However, a new rule iterates that the people's decisions on the use of their personal information must be well informed, and through a separate affirmative action by the user, where he/she has a choice to use the AI system without compromising his privacy.

Moreover, this study presents a set of principal actions that must be addressed to ensure the AI system is indeed ethical and possess the characteristics of an ethical AI:

2.1.2 Principal for Developing the Ethical AI [8]

- Impact assessments: An ethical AI should measure its impact on the public, through questions such as what are the direct and indirect impacts of the AI system on different ethnicities, communities, or political groups? As an Example, Westpac bank in Australia used facial recognition for recognizing the mood of its employees [18]. An interesting view of this situation is delivered in this study, which highlights the value of assessing whether the AI system is designed to benefit the employees or maximize the profit. They argue that it is the machine's purpose to serve the human, not the other way around.

- Risk assessments: For every AI system, risk assessments must be conducted to analyse whether or not it may create or induce any threat to the public. A hypothetical example is, when an AI system is designed to find the optimal route for oil pipelines and is biased toward indigenous people or in general minorities, such a system can cause great discomfort for them, which can also have political costs for the government.

- Consultation: If an AI system affects a community, they should have a voice to express their concern or opinion about the ethical issues that must be addressed, even within the development stage of the system.

- Recourse mechanism: There should be a mechanism set in place so that, if a person or community is negatively affected by an AI system, they can formally complain about the system or its developer.

- Mechanism for monitoring and improvement: The AI system should regularly be monitored and held accountable so that its relevance, accuracy, and decisions be in-line with the ethics, and needs of the society at the current time.

  An example of this issue is mentioned about the use of facial recognition in the United Kingdom, by the police department, in the London metro area. Based on the report by Big Brother Watch [19], most of the result of the facial recognition system was proved to be inaccurate, and this method is lead to a significantly small sample size of arrests. Here, they argue, this significant difference will result in public distrust, and [19] suggest that maybe a solution is that, before questioning and confronting the suspect, a human may assist the AI system to ensure that identification is correct.

  Several important factors exist when considering the monitoring and improvement of the AI systems, in [20], some of the basic ones are signified. In-house development, which alternatively, [21] refers to it as a secure enclave, is one of these important factors. It means limiting the accessibility of the data to the attackers, by mandating rules that prevent the leak of client's private data or information about the AI structure to the public. For example, [21] mentions that the rapid growth in the utilization of the public cloud is an important aspect of the increasing number of cyber-threats against the AI systems. In [21], the solution is regarded as promoting the secure execution environment. Software Guard Extensions (SGX) designed by intel is an example of a hardware-enforced isolated environment, which protects the code inside the SGX from the compromised operating systems. Another important factor in cyber-security is referred to as Know-How, which means how much the adversaries are aware of the structures of a certain system. This would define what type of attack they can perform on a certain AI system, which also determines the ability for monitoring and protective measures that can be considered for the AI system.

Another factor that is mentioned in [21], is the cloud-edge systems. It means increasing the number of data centers and making them closer to the user. Right now, a few companies such as Google and Amazon manage the major data centers for today's information distribution, which means due to the centralized nature of the information distribution, the data travel through many routers, which jeopardize security and safety of the AI systems.

- Guideline for best practices: Developers should be provided with a unified, and transparent guidelines, this guideline can be considered as a checkbox that developers will follow for ensuring that their product uses a representative data, its security aspects are ensured, and will not cause any harm.

- Industry standards: Providing the developers with certification and standards is essential, and can confirm if an AI system follows the ethical guidelines, this may also induce competitiveness among the developers.

  In this regard, the Office of Australian Information Commissioner has called for developing a standard for AI developers under the title of "How standards can increase organizational accountability" [22].

  Besides we can argue that this standard should take the cybersecurity aspect of the AI system into account and make sure that the safety and privacy of users are the main priority.
- Internal/External view: Using an unbiased selection of professionals to review the AI system, and examine it with consideration of the ethical principals. Here, the external views can be made possible through assigned unbiased investigators, that are responsible for measuring how much an AI system follows the industry standards, or developers should be encouraged to reach out to the public and use their views. Besides, internal views can be obtained by hiring and taking opinions from diverse unbiased professionals.

- Collaboration: Using the academia alongside industry to develop "ethical by design" AI. This collaboration tries to answer the integration of ethics into the AI systems.

Here, as a result, the "guideline for best practices" can persuade the developers to work with academia, so they go through another accountability layer. In this study, the presented framework implies that without laws that enforce transparency by the developers, getting feedback from the public, and setting up a standard, one cannot monitor the potential outcome of an AI system on the society.

European Union Agency for Cyber Security (ENISA) is one of the established organizations working through the standards for cybersecurity certifications. Some of the tasks of this Agency are introduced as the following [23]:

- "Cyber-hygienes" must be promoted, meaning that cyber threats will not solely affect the technology but also the citizens.
- Raising awareness of citizens and business owners against cyber-attacks.
- Decreasing third-party dependency of the AI systems.
- Performing regular analysis and maintaining a "market observatory".
- Promoting cyber-security certification.
- Promoting international collaboration.
- Promoting regular consulting with standardization organizations.
- Thinking of flexible solutions, so the industry remains ahead of the threats.
- Promoting a high level of safety assurance for Information and Communication Technology (ICT) products.

With regards to the implementation of what ENISA is advertising, in a similar viewpoint to this report, a study is published in the journal of nature in December 2019 [20]. In the mentioned paper, three elements are introduced, which makes the AI system more reliable. One key element is adversarial training, which will be discussed in this report. In this paper, adversarial training is advertised as a mandated rule for the AI system to improve its performance. The second element is referred to as in-house development. This study argues, in the development stages of an AI system, the requirement of using cloud systems due

to receiving support from third parties makes them vulnerable to many possible types of attacks, and an in-house development would prevent such threats. Here, the key argument is decreasing the space of possible threats. The third element is parallel and dynamic monitoring. They argue that the developers should construct a "clone system" as the observer system, and they explain that such a system should not be considered as the "digital twin" [24] of the main AI system. Furthermore, they demonstrate a set of reasoning and design criteria for the clone structure that would deliver their desired monitoring performance. These characteristics are explained as the following [20]:

- Clone system is the same AI system but in a controlled condition.
- Clone system will go through adversarial exercise and situations where the original system was not assessed.
- Threshold divergence between the two AI system's classified outcomes will be considered as an attack.

In this report we want to go through elements that together would form a transparent AI. These elements are directly related to deficiencies of the AI systems in several categories. First, we discuss these categories one by one and will recommend a accountability model based on these categories.

2.2 Data Governance and Privacy Issues Related to an Ethical by Design AI

In the previous section, privacy is introduced as one of the core principles of ethical AI. Here, we address how data governance is essential to the privacy of the users. Here, we address this problem through the review of important definitions and issues presented by [25].

Regarding consumer data, [25] mentions three current methods for which data is being collected about consumers by companies, based on this study, they are regarded as declared, observed, and generated data. Declared data is a consented directly requested data is that the user agrees to share with a website or application, which can also be regarded as active data. Observed data refers to data, obtained through tracking user's social media interaction and their

research preferences. Generated data is the data that is obtained by analysing the information gathered about a user. The following table displays some of the means of data collection.

Some Means of data collection [25, 26, 27]			
Data	Active data	Passive data	Description
User's name	X		
User's address	X		
User's preferences	X		
Cookies	X		Collected through interaction of a user with a web-site, third party companies can buy and sell this info to other companies.
Device information and tracking		X	User agrees when he install an application.
Facial recognition		X	Utilization of biometric software that identifies person through their image
Fingerprinting		X	Companies can use specific information that is gathered about the user as the specific identifier for that user.
Payment cards		X	They can track member's transaction history.
Cross device tracking		X	They tack users and can pinpoint them to devices like as mobile and computer

Table 1: Types of passive data collection by [25]

In [22, 25], they argue that third-party companies, buy people's data, which is collected via passive and active means of data collection, through the companies which the user agreed to share them with, and sell them to other groups and organizations for alternative purposes.

To further address this topic, we have to clear what does personal data means. In [27], they define the name, email address, home address, phone number, and work history as directly personal data. In [25] and [27], they agree that some data can be indirectly personal, which also enables the identification of the user. Also, based on [27], some personal data are more sensitive than others, such as health, religious beliefs, political views, and race. Therefore, a strong definition of what is considered personal data is needed. Based on GDPR, data is personally identifiable if it can be linked to a personal computer or device. This is a reason why, in the new rules imposed by the European Commission on data regulation [17], a person must take an informed decision on which data can be used by an AI system, also, utilization of the data by AI system, should not be essential for the permission of the application usage. Under General Data protection regulation the following are some of the basic rights of the users with regards to an AI system
[17]:

- (i) Data protection by default: Companies should not decide whether a user's data should be accessible by public or other companies and third parties, this should happen through an informed decision by the user, and if a user's data is leaked, which may cause damage toward the user, the company should be held accountable.

- (ii) The right to ask: A user must be able to ask about the information type and its processing purpose by the company.

- (iii) The right to object: Users should be able to object if they don't want a company to process certain information about them.

- (iv) The right to access: Users should be able to access all the information that is kept about them by a company.

- (v) The right to be protected and informed: If personal data is leaked, the user has the right to become informed about it.

- (vi) The right to be forgotten: Users should have the right to request for deleting all of their data. In a situation where this information concerns public interest, the benefit of the majority prevails. In 8, they argue that

this right is a challenging issue, since the data may already be integrated with the structure of an AI system, therefore they argue that Australia may wait to see how EU can integrate this law into the AI system.

In an example of this view, which is mentioned in [21], published by the University of Berkeley, they indicate the importance of sharing information on matters that concerns public health. They describe a scenario in which hospitals are likely to share private health information of patients between themselves to improve the response of the government to a situation such as an epidemic. For example, in a situation where a vaccine is found for a type of flu, which regions should be prioritized to receive the aid first.

2.3 Data Selection and Biases

In [8] it is implied that how ethical and policy issues in data selection can lead to unethical and untransparent practices by companies. In [28], a study is reviewed that talks about the price of online tutoring by Princeton university with regards to the ZIP code. This study shows Asian living areas in the U.S are charged up to two times higher prices for online tutoring than the other areas, also this study unveiled that the wealthy neighborhoods of Washington DC were also charged higher than average by this university. The officials of the Princeton university denied such discrimination, though, the available tools are not simply strong enough to enforce transparency. Therefore, another issue lies in the use of deep learning algorithms, which are not explainable, if the law could enforce an answer from the university and the methodology that their AI system utilized was transparent, an investigator could examine the decision for online tutoring prices with regards to neighborhood plus the explanation for that decision. This type of investigation may reveal wrongdoing and lack of ethic in many cases and help to regulate AI developers. This matter will be more investigated in the context of explainable AI through the next sections.

2.4 Transparency of Methodology and Explainable AI Design

The issues regarding the AI systems necessitate the need for constraints that should be considered in the decision-making process and data preparation of an AI system. As mentioned, ethics, privacy, security, and fairness are characteristics

of an AI method that is transparent. One of the available tools for designing such systems is explainable AI/ interpretable AI. In this section, the main challenge is to identify the elements that will together form a truly explainable AI. To avoid confusion, we will use the term ethical AI for the system that has all the characteristics. The explainable AI method is just the first of many elements for achieving such an AI system. As mentioned, an interpretable AI system is an improvement that will help to facilitate debugging, increase understanding of the AI by the user, decrease subconscious biases, and brings trust toward the performance of the AI system. An AI system that has this characteristic is called an interpretable/ explainable AI in terms of methodology. First, we try to describe the available explainable AI methodologies and discuss what it means for an AI system to be explainable in terms of methodology, then we move on to highlight the other important aspects of a safe and trustable AI.

2.4.1 Explainable AI

Explainable AI methods that analyse the AI techniques based on the type of their interpretability are classified into the subclasses of post-hoc and intrinsic. Intrinsic interpretability means that the AI method is interpretable due to its simple structure, while post-hoc interpretability means that interpretability is achieved after the AI model training. Another important definition provided in this study highlights the degree and scope of information provided by an explainable AI method. Based on this definition, an explainable AI methodology can either be globally interpretable or locally interpretable. An explainable AI model is said to be globally interpretable when it follows entire reasoning leading to all different possible outcomes. If the explainable AI model only explains the decision for a specific situation, then this is considered as local interpretability. This paper also categorizes explainability techniques based on their applicability to different AI systems into model-specific and model-agnostics. While model-specific methods are only applicable to a specific type or class of algorithm, model agnostic methods are usually post-hoc techniques that are not tied to a specific class of machine learning algorithms. In [29], a categorization of explainable AI techniques based on their underlying structure is provided, in the following a brief description of these techniques is provided:

1- Surrogate model

It is a simple model that is trained on the prediction of the black-box model to explain its complex structure. Decision trees, LIME(Local Interpretable Model agnostic Explanation), and the linear model are examples of these types of models. There is no theoretical guarantee that a surrogate model can represent a complex model.

2- Partial Dependence Plot(PDP)

The partial dependence plot provides a visualized explanation that highlights how predictions depend on one or more input variables.

3- Individual Conditional Explanation

It is an extension of the partial dependence plot, where it explains in terms of a graph, how the prediction changes when a feature of interest changes. This method can only provide a meaningful explanation for one feature, by considering two or more features, the explanation plot will be a surface or a hyperplane and it cannot provide a meaningful explanation for the user.

4- Rule Extraction

It provides a symbolic human interpretable explanation that approximates the decision-making process of the artificial neural network. Rule extraction methods are generally categorized into the decompositional approach where the network is broken down into smaller individual parts, the pedagogical model agnostic approach where the target is the function computed by the network and its input features are the network's input features, and eclectic approach where it uses the knowledge about the weight vectors in the neural network model to complement its symbolic explanation [30].

5- Model distillation

It is a model compression approach between the deep and shallow network, where a smaller model is trained to represent the larger model. This is done by

transferring the knowledge from the larger model to a smaller model through minimizing a loss function.

6- Sensitivity analysis

It is referred to the analysis made on the effect of input or weight perturbation on the concluded output by an AI system.

7- Layer-wise Relevance Propagation (LRP)

It is a technique where it propagates the outcome backward using a set of locally designed propagation rules [31].

8- Prototypes and criticism

This model includes two parts called as prototype and criticism. Prototype is an example-based explanation, where representative data instances are selected to bring an understandable view to the complex data distribution. Criticism is an example-based explanation of data instances that are not covered by the selected prototypes.

9- Counter factual explanation

Proposes a minimum alternate explanation that would lead to a different outcome in an AI system. In the following a table representation of the aforementioned techniques is provided [29], which divides the available studies in the literature based on the provided definitions.

Furthermore, these methodologies help us to move on from the black-box model to something clearer, which is also referred to as a glass-box [32]. Based on [32], these techniques can be used to recognize racial, gender, intrusion, or locational bias in the systems. They can be regarded as one of many criteria that we can set for defining a secure system. An important question is when the output of an AI system is not explainable, can we still consider it a safe system? This is an issue of risk assessment, where there is a trade-off between vital aspects of the task and ethical issues of the AI system. We suggest that it should be assessed whether having such an AI system is safe for the public or not.

Explainable AI (XAI) methodologies have the potential to have a critical part for numerous applications in the industry, defense, civilian, or medicine. For example, DARPA (Defense Advanced Research Project Agency) developed a model, based on RISE (Randomized Input Sampling for Explanation) for identifying solar farms through satellite image, they have successfully recognized important features in the input of a neural network program that highlights why the neural network recognizes that a particular image belongs to the solar farm and why the image may display a different structure [33].

Another example provided by this agency is based on the idea of the network dissection, which means they identified what neurons inside the hidden layers of a neural network program, are responsible for detecting a place such as a shopping mall, a park, or a solar farm [33].

In [33], deep reinforcement learning also has been mentioned as an alternative explanation methodology for a black-box model with complex neurons. Here, the idea is that to replace neuron operations inside the hidden layers, with a simple gradient-based dynamic to be able to simplify the complex model. As mentioned earlier simplicity boosts explainability, as it would enable us to analyze these complex structures with well-defined methods [33, 34].

In [35], an example of explainable AI for biological applications is mentioned. The support vector machine is a classification and a regression method that is used for a variety of data analysis and has long been used in the field of biology. As an example, in [35], they investigate an AI system which has to detect a specific biological signal in a sequence. However, the output provides no meaningful explanation for its resulting decision. Their methodology was able to produce a biologically relevant explanation for the studied machine learning program.

In [36], a study categorizes the opportunity and challenges that can deeplearning and interpretable AI bring to the fields of medicine and biology. In this regard, disease and patient categorization for classifying cancer patients, predicting gene targets of micro RNAs, or electronic text mining of medical health records aim to retrieve relevant information that are beneficial to public health.

In [37], an example of XAI in the medical field is investigated. This study examines the development of an XAI for a deep neural network that is designed for the prediction of the health status of the patients in the intensive care unit (ICU) of the hospitals. The subjects of this study, which was done by the

researcher of the University of Southern California, are the patients with acute lung injuries. Based on this paper, the AI system uses the health data from sensor signals such as temporal variables, ventilator settings, and blood gas value that are provided by ICU machines, in addition to the medical records of the patients, to decide whether the patients' health will likely elevate/worsen over a period of time. Here, the AI system predicts when the hospital will have a ventilator-free day and how the pattern of the mortality will change over time. When a hazardous event such as the COVID-19 virus is threatening human lives, it is evident why such an AI system and their respective explainable AI will play a critical role in our life.

Another interesting example is the implementation of an interpretable visual saliency method [33, 38] where the AI system is responsible for deciding the movement of the vehicle, and XAI method is responsible for rationalizing the movement with a textual explanation. This methodology generates a heat map, which takes into account variables such as traffic light and explains why a vehicle is slowing down. An interesting perspective that can be concluded is to consider investments for developing detectors or observers that operate based on the explainable AI. Imagine a scenario where a deep neural network and control system is responsible for the movement of an autonomous vehicle, in this situation, if such an explainable AI-based detection system, provides an explanation based on its received variable, that the vehicle is accelerating while the light is red or the front vehicle is slowed down, then the system may recognize the problem with the decision of the AI system. However, in a situation where the resulting explanation for AI system commands is not explainable through measures like as explainable AI, maybe we should consider that system unsafe and unauthorized to operate.

As mentioned earlier the implementation of explainable AI will provide a piece of information that is human-understandable. Here we argue that one of our focus should be to encourage or even train developers for such programs. Imagine for the mentioned case of recruitment AI, there was a parallel interpretation model that could provide an explanation for the Houston teacher, highlighting the reasons that cost his job. Proceeding with the legal course would be much easier if there was a decision that was assessable.

Categorization of studies related to the explainable AI By [29]				
Category of technique	References	Global / Local	Intrinsic / Post-hoc	Model agnostic/Model specific
Rule lists	[39][40][41][42][43]	G	I	SP
Decision Trees	[44] [45] [46] [47] [40]	G	I	SP
Shapely explanation	[48]	L	H	AG
LIME	[49][50] [51][52]	L	H	AG
Saliency map	[53][54][55][56][57][58][59]	L	H	AG
Rule extraction	[60][61][62] [63][64][65]	G/L	H	AG
Decomposition	[66][67][68]	L	H	AG
Partial depen dence plot	[69] [70][71]	G/L	H	AG
Feature impor-tance	[72][73][74]	G/L	H	AG
Prototype and criticism	[75][76][77][78]	G/L	H	AG
layer wise Relevance Propagation (LRP)	[79]	G/L	H	AG
Counter factual explanation	[80]	L	H	AG
Model distillation	[81][82][83][84][85][86]	G	H	AG
Sensitive analysis	[87][88]	G/L	H	AG

Individual conditional explanation	[89][74]	L	H	AG
Surrogate models	[49][90][91]	G/L	H	AG
Activation maximization	[92][93]	G	H	AG

Table 2: Categorization of explainable AI techniques by [29]. In this table "G"21 refers to global, "L" refers to local, "H" stands for post-hoc, "I" stands for intrinsic, "AG" points at the model agnostic methods and SP refers to the model specific methods.

Furthermore, in important question arises: What are deficiencies that each of these field faces which might delay the integration of AI, or XAI in their respective domain:

Deficiencies of AI in Healthcare

For better integration of AI systems in healthcare, we must first identify what preliminary measures are required to prevent data bias, increase accuracy, and elevate the reliability of these systems when the public needs them the most. Based on [94], the first step is to increase the capacity of the data centers for local hospitals. This paper mentions that as of now, many local hospitals cannot store large amounts of electronic health records over a long period of time. As a consequence, lack of comprehensive data is detrimental to the accuracy and reliability of the AI system. The second element is attention to bias in data. For example this study mentions that minorities might not be accurately represented by the AI system that is trained for the people with different ethnicities. As mentioned and also highlighted by the aforementioned research, lack of explainable AI in healthcare systems is another aspect that is currently missing, where a doctor might need to be aware of the logic behind the decision of the AI system before any final decision has been made. In [94] a study is mentioned, in which the AI system is responsible for identifying patient that have a higher risk for pneumonia, in this study patients who were suffering from asthma, were categorized as a lower risk to be endangered by pneumonia, while in reality such patients are at higher risk. This study also signifies the need for XAI in health care.

Based on [95], an important factor is to be able to break the private contracts between patients and hospitals on the confidentiality of the information, in a situation where it is necessary to have an analysis on a bigger scale, such as epidemics. Other important factors, which [95] recognizes, include enhancing

technical knowledge, removal of biases, and explainability. Moreover, there is a consensus on the transparency of the AI system by using methodologies that are engaging with their respective users. This study argues that experts which develop the AI systems must be aware of detrimental biases. This matter can be resolved with proper training in clinical education, hiring the qualified expert, and normalizing this point of view amongst the expert and policymakers. Alleviating the deficiencies through the experts' help is the key element that this paper mentions for improving the AI systems. For example in every field, there are important questions that aim to examine if their respective AI system can be an answer to them. For example, this paper mentions the area of radiology in which the Royal College of Radiologist in the united states defined an AI framework, which the developers may consider during the design of their product. Another aspect is to make the radiologist familiar with the bias and training procedure of the AI system. This will enhance the quality of recommendation by them, which is directly in line with the deficiencies of the AI systems. Another aspect is advertising, this is the key element for bringing public trust toward an explainable AI system. Many of the current public opinion about the AI is system is far from reality. It can be categorized into skeptics or people that think an AI is an answer to everything. To be able to answer the skeptics, we must advertise the explainable AI and its benefit to the scientific community, so more people be encouraged to be engaged in research and development in this field, and also more scientist be able to trust explainable AI decision in a critical field such as healthcare.

2.4.2 Adversarial Examples in AI

In terms of the capabilities of the adversaries, the adversarial attacks are regarded as attacks that are injected into the AI system during the training stage or during the testing stage of these systems [96] that are also referred to as the adversarial examples. Adversarial example intends to deceive the AI system and cause misclassification.

Based on [96], three types of attacks are possible during the training stage: data injection, logic corruption, and data modification. During the testing stage, this paper categorizes the attacker's capabilities to the two-class of the white-box or the black-box attacks.

In white-box attacks, the defender assumes that the adversary has total knowledge of the targeted AI system. Based on [96], there are two general ways for designing such attacks: sensitivity estimation and perturbation selection. They explain sensitivity selection as finding the direction of input, which by applying minimum disturbance, the intruder can cause the misclassification of

the output. Based on [96], perturbation selection is defined as finding a disturbance for the input that results in the desired misclassification.

In black-box attacks, the assumption is that the adversary is not aware of the AI system's internal structure [96], and the access is limited to the input and outputs. Based on [96], the black-box attacks are sub-categorized into non-adaptive, adaptive, and strict black-box attacks. This categorization is based on the adversaries' capability to analyse the information about past inputs/outputs data, which means the attacker can interpret and analyse the original model, and use his/her model/analogy of the AI system to generate attacks. A real-world example of black-box attacks is investigated in [97]. This paper discusses how a deep learning method can be used to generate an attack against the AI system. To prove how deep learning can work against deep learning they conducted attacks on MetaMind, Amazon, and Google. In an example provided by this paper, they illustrate one of their attacks, when they targeted a DNN benchmark designed to recognize the traffic signs, and forcing it to misclassify the input data.

Furthermore, there are techniques in the literature that are specifically designed for this purpose. In [98] a study is conducted that investigates the adversarial image generation. It introduces three methods of fast, basic iterative, and iterative least likely for this purpose. Fast and basic iterative are attacks that add a gradient-based designed perturbation into the pixel values, while least likely is a more sophisticated version of the adversarial attack. An example of the least likely is when image recognition classier mistakes a dog as a plane [98]. The following table demonstrates some of the known attack that are mentioned in [96]:

Adversarial Examples by [96]		
Attack's name	WB/BB	Description
L-BFGS	WB	Adding a gradient-based designed perturbation into the input value.
Fast Gradient Sign Method (FGSM)	WB	Adding a gradient-based designed perturbation into the input values.
One-Step Target Class Method	WB	Another type of gradient-based method.
Basic Iterative Method	WB	Another type of gradient-based method
Iterative least likely-class method	WB	A probability-based version of the gradientbased method.
Jacobian Attack Saliency Map	WB	Applying the maximum perturbation to targeted few inputs
One Pixel Attack	WB	Changing only one pixel for attacks against facial recognition systems.
Deep Fool	WB	A simple method to fool the neural network algorithms
Houdini	WB	Using gradient based information to target machine learning algorithms designed for application such as speech recognition or facial recognition systems.
Model Inversion	BB	Using machine learning to recover the input data.
Model Extraction	BB	Building a clone model with a similar structure as the targeted model.
UsingParallel Model	BB	Devising a clone model based on the output generated by the adversary input.

Table 3: Different types of white-box and black-box adversarial examples reviewed by [99]. Here, WB means white-box and BB means black-box, sensitivity estimation is denoted by SE and perturbation selection is denoted by PS

.

Furthermore, in [99] a study developed on the security of machine learning. Based on this work, in order for the user to trust an AI system, first, we have to

study this through the viewpoint of the attacker to identify the intent of an attack design. This study discusses the classification of attack types against AI systems that would cover applications such as spam filtering, virus and worm detection, and intrusion detection. Based on this study, malicious inputs based on severity can be classified into two classes of harmful and benign that can have a false positive or false negative impact on the AI system. False-positive is the same as a false attack where normal input is detected as an attack, and false negative is when the attack situation is considered normal. Based on the false positive or false negative, this work elaborates the idea further to categorize the malicious input that can have such an impact on the system. *Table.4* demonstrates this categorization.

Adversarial goals by [99]			
Categories of attacks	Purpose	Impact on availability /false positive	Impact of integrity/ false negative
Targeted attack	Focuses on a specific output	X	X
Causative attack (poisoning attack)	Takingcontrol of training data	X	X
Indiscriminate attack	Targeting a variety of the outputs	X	X
Exploratory attack	Taking advantage of misclassification	X	X

Table 4: Adversarial goals by [99]

Moreover, studying adversarial examples and intent of the attacker simply can be considered as a guideline to encourage developers to become more proactive.

This aim proceeds with analyzing the weakness of the AI system before the occurrence of an attack. Identifying the possible intent of an attacker is the key element of recognizing the weaknesses of an AI system. Another element is studying the design methodologies that exist for the development of an attack. What needs to be done is to decrease the level of uncertainties that exists in the AI system. Table 3 mentions several types of white-box or black-box attack, a developer's job is to address these attack and provides assurance which scales what level of security their product provides against these attacks. However, the criterion is recommended to be universal and we will mention it as part of key consideration that must be enforced by the policymakers, technical researchers, and inspectors that set the guidelines for a safe and trustable AI.

Several important questions arise from the preliminary investigation on the concept of adversarial examples in AI, which we aim to answer by exploring the literature:

- What is the nature of methodologies in the literature for designing an adversarial attack?
- Since the achievable interpretability of an AI system does not automatically result in detectability, how can we use adversarial attack combined with interpretable output to improve the notion of detectability?
- Is there exist a criterion on how much an AI system is robust against adversarial attacks?
- How can we identify and remove adversarial input data from the AI system?

2.4.3 Defence Strategies Against Adversarial Attacks

Based on the available strategies in the literature against adversarial attacks, [96] categorizes the defense strategies into modifying data, modifying models, and using auxiliary tools. Methodologies classified as modifying the data are concerned with the training stage or the testing stage of the data sets. Modifying Model strategies recognize that the solution against adversarial attacks lies in increasing the adversarial resistance of the AI system. Additionally, the auxiliary tool is an artificial tool that assists or defends the AI systems. Each of these strategies addresses a specific type or a range of adversarial attacks against the AI systems. According to [96], methodologies based on modifying data

include adversarial training, gradient hiding, blocking the transferability, data compression, and data regularization. In this regard, modifying the models encompass regularization, defense distillation, feature squeezing, Deep Contractive Network (DCN), mask defense, and Parseval network. And lastly, methodologies using auxiliary tools include MagNet, high-level representation guided denoiser (HGD), and Defence-GAN. As an example, gradient hiding is introduced as a type of defense mechanism which is effective against gradient-based attacks. However, MagNet is an auxiliary tool, that operates as a detector for the main AI system, which can be utilized as an active defense against any adversarial intrusion. This implies that the developer and any potential standard provider must decide which strategy is best suited, based on the application area of an AI system. Based on this paper, we can also conclude that most of the methodologies proposed against the adversarial examples can be categorized into passive or active methods. They can also be regarded as attack specific or as attack agnostics.

In this report, one of the primary objectives is to address the cyber-threat besides the discriminative issues regarding the AI systems. Through-out this report, we try to present the adversarial threat and their solution in an organized manner so that collectively along with the other measures they can reflect a safe and unbias AI system. In light of this approach, *Table.5* displays the existing strategies in regard to cyber-threats in the literature, which are categorized by [96]:

Defense strategies against adversarial examples [96]	
General types	methodology
Modifying Data	Adversarial Training Data Compression Gradient Hiding Blocking the Transferability Data Compression Data Randomization
Modifying Mod-els	Regularization Defensive distillation Feature Squeezing Deep Contractive Network (DCN) Mask Defence Parseval Networks

Using Tools	Auxiliary	Defense-GAN MagNet High-level Representation Guided Denoiser (HGD)

Table 5: Defense strategies against adversarial examples by [96]

Here, we argue that besides the knowledge of possible adversarial examples, it is crucial for an AI developer to be aware of the defense strategies against such attacks. Basically, we argue that developers must recognize the holes in their AI system as well as preventive measures against such vulnerabilities.

Until now we present an introductory the explainable AI methods, *Table.2* presents an overview of available techniques for the development of an AI without bias [29]. These techniques overall, present a simplified explanation that can be interpreted by the human user. We also discussed adversarial examples and the available defense mechanism against them in the AI systems. One of the widely discussed defense methodology is the adversarial AI. But, how can we be sure that an explainable AI or an AI system does not have any loophole? Reverse explanatory techniques that are developed for the AI system can address this problem. They are called as adversarial machine learning [10].

2.4.4 From Adversarial AI to Explainable AI

In this category, the designed adversarial attack by the defender attacks the AI interface through its vulnerabilities e.g. where the data is not available. This technique first finds adversarial samples that are within the valid entries of input space, such that it maximizes the impact of the attack on the system. When the interface misclassifies the attack as a normal event, the adversarial machine learning tries to find the reasoning behind it [100].

Based on numerous works that are presented in *Table.2*, the explainable AI is suggested as a defense mechanism against the adversary attacks. However, in [100], the idea is expanded and approaches the possibility of an adversarial attack against the explainable AI. They utilized linear and multilayer perceptron for explaining the adversarial data against the explainable AI.

An illustrative example of adversarial data can be found in [101], where an adversarial data misleads a VQA (Visual Question and Answering system). Here, the VQA system incorrectly recognizes a traffic light red, while it was green. Another example can be found in [102], where the attack detectability is investigated for the particular case of Face Morphing attacks. Face morphing attack manipulates the result of the face recognition systems to be the face multiple identities, which criminal individuals use in instances such as border crossing. In this regard, a practical study conducted by [103], that demonstrates how wearing an adversarial purpose designed glass can help the criminal evade the face recognition system.

In [10], similarly to [100], the problem of adversarial AI is discussed. Here, this problem is elaborated in the context of repeated games for the Bayesian classifier in the framework of the AI system. Consider a situation where an adversarial methodology is developed that would find the data for which the optimal attacks on the original classifier is possible, the outcome is an improved classifier. A repeated game is when the optimal strategy for finding the adversarial samples, changes indefinitely by taking into account that the adversary has the knowledge of the updated classifier. Aside from the repeated game, this paper introduces different challenges that must be considered to become a step closer to an AI system that is self-aware of its own vulnerabilities. Sub-optimal strategies, generalization to the other classifiers, multiple adversaries, incomplete information, and false attack are the important topic of investigation that are mentioned in this paper.

Sub-optimal strategies become necessary when the desired optimal strategy has a high computational cost. Incomplete information considers a situation where adversaries and classifiers do not possess the perfect knowledge of each other. Multiple adversaries refer to situations where the AI-system is under attack from numerous sources e.g multiple spammers. Lastly, the false attack means that the goal of the adversary is to find healthy data that the AI-system would classify as an attack.

Here, we discussed the adversarial AI, and we emphasize that it should be considered as a necessary element in designing an AI system by the developers. This methodology helps to boost the security of the AI system. In addition, it is also important for developers to become familiar with adversarial examples and other defense strategies against them. In the next two sections, we will briefly

describe these pieces of security in dealing with AI systems that can help us protect the stakeholders' interests.

2.4.5 Discussions Regarding the Detectability

In the following, some of the studies that consider the notion of detectability for the AI systems are addressed. In [104] the authors try to explain the adversarial attacks (also referred to as adversarial examples). An important conclusion of this paper is that the over simplicity brings vulnerability. They raise the question about a phenomenon that "why multiple classifiers, misclassify a data in the same way?". In another word, why would different nonlinear classifiers, treat an adversarial example in the same way. They elaborate that this phenomenon occurs because they belong in a precise location of the output space. They hypothesize that adversarial examples are the result of models being too linear instead of nonlinear. They explain this by suggesting that the weight update of most classifiers follows the same reasoning. As a result, most of the trained classifiers by the neural networks will produce similar weights. They argue that linear classifiers are more vulnerable than nonlinear classifiers. However, a trade-off would emerge as [29] addresses that the linear classifiers are better explainable. Although they refer to a study done on the property of neural networks, one should be cautious to generalize that most classifiers output is the same for adversarial data. However, this topic is very interesting to investigate, and actually, it may be able to enhance the security of AI systems. Moreover, more studies should be developed to test specific adversarial examples against different AI systems to study their interaction in a better, organized framework.

In [105] another important issue is addressed that would bring the light into the detectability of attack. This paper aims to find the relation between the detectability and the strength of an attack. The paper argues that the boundary set on the input data plays an important role in the detectability. They say that if the boundary of acceptable data by explainable AI is high then it would be easier to detect the attack as the outcome of a not well-designed attack would be an unrealistic output and an interpretable AI can detect it. However, what they are arguing would not limit an attacker that has knowledge about the input of the system and actually, the attacker will be provided with a larger space of possible attacks. To overcome the aforementioned problem, they propose an outlier detection strategy to remove the bad data from the trained AI. Here, this is accomplished with the help of a distance threshold that measures the distance of incoming data from its nearest data point. They provide simple reasoning that a causative attack will not be possible if the causative data is removed.

2.4.6 Security Certification Methods for an AI System

In this report, we aim to review studies conducted on the measures for the security of the AI system. Until now we could not find any universal measures on security, but some elaboration of the concept can be found by going through the literature. In [106] a study on the security of AI systems against adversarial attack is provided. They discuss the notion of adversarial robustness for the AI system. The idea here is that: "how can an AI system be trained, so that it is robust to adversarial data". They developed a measure for quantifying the security of the AI system against adversarial attacks, through a min-max optimization problem. They claim that this notion although non-convex or non-concave is tractable. However, the contribution seems to only cover one type of bounded attack and it does not provide a universally acceptable index.

The vulnerability of the AI system to adversarial examples opens up the discussion about the robustness of these systems to such attacks. In the mentioned study on the security of the AI system, the notion of adversarial robustness is defined for one type of first-order gradient-based attack, and with [107] we try to explore further into this topic.

In [107], they presented a criterion referred to as AI^2, which can certify the robustness and safety of an AI system with abstract interpretation. The ability of this method is tested against feedforward and convolutional networks. First, we have to become familiar with the language used here for defining the adversarial robustness. In this paper, the properties of a tool that can verify the robustness of an AI system are precision and scalability. It means that this tool must be able to analyze the output of an AI system over large sets of data while providing an acceptable precision. As an example, consider a picture referred to as C, they want to demonstrate if we filter this image with a specific intensity located in a specified region, are we still able to identify that image as its class, in another word, what is the region of perturbation for the examined class for which the AI system can handle the variation, and identify the accurate class. The concept that they use for this purpose is referred to as abstract interpretation. Here, abstract Interpretation is defined as "Theory that gives a finite approximation of potentially infinite sets of behavior". It provides the ability to analyse the neural network over an abstract domain. This means we can over-approximate the results of a neural network over shapes such as zonotope. In this regard, DeepZ [11] is another tool that uses abstract interpretation for certifying neural networks. Such methodologies will provide the developers with a tool that can certify the response of their output to adversarial examples, which can be considered as another aspect to the standards in designing an ethical AI.

In the next section, we try to argue about the critical elements that are discussed here as part of a certification process that can be considered for the AI systems.

2.5 Technical Certification of an AI System

Transparency of methodology and using methodologies such as explainable AI, security certification, and adversarial AI means that we should not be focusing on setting the standard based on the trust but scalable criteria. There has to be an assessable mechanism that can define this trust. The proposed argument of the [20] for parallel and dynamic monitoring, while interesting does not answer the legal trouble that may occur due to the unexplainable nature of the AI systems. For example, if an AI system damages its client's interests, the proposed observer of this paper will not provide a sensible assessment that will help to settle legal cases as it operates solely based on a threshold checking.

Here as introduced in section 3, we argue that such monitoring may be established through the utilization of an explainable AI system. If an AI system is unable to explain its decision regarding a particular output, in a human-understandable way, it should be considered as one of the red flags that can be caused by the lack of proper training or a cyber attack. Therefore, we argue that an observer / AI system based on adversarial training and explainable AI should complement each other and provide unified reasoning that would help indicate attacks, as well as explain the resulting decision of an AI system.

In the following, *Table.6* is presented, which demonstrates this report's views of the methodologies presented in the literature, and what impact they can have on resolving the legal, as well as the cyber-security issues, which are/ and, will be present by the growing integration of artificial intelligence. For example, in this table, adversarial AI methods are mentioned, which usually enhance the security of the AI system to cyber-attacks in a passive manner. On the other hand, active monitoring methods evaluate each resulting decision of the AI system and can determine whether the AI system is under attack. If combined active monitoring based on the explainable AI and Adversarial AI are established, in such a way that the outputs of an AI system are consist of an explainable result, the resulting decision, and a decision provided by the monitoring system, these decisions can be found crucial, in settling legal issues as well as improving the cybersecurity of AI systems.

Transparency of the methodology					
Methods	Passive/Active	Improves AI Security	Transparency of decision	Improves the AI Method	Evaluates the AI security
Interpretable/Explainable AI	Passive	X	X		
Adversarial AI	Passive/Active	X		X	
Security certification methods	Passive				X
Parallel and dynamic monitoring based on Adversarial AI	Active	X			X
Monitoring based on explainable AI	Active	X	X		X
A combined exAI and adversarial AI technique	Active	X	X		X

Table 6: Categorization of methodologies that can be used for improving transparency and decreasing bias in the AI system.

Furthermore, besides the ethical regulations and pre-implementation of preventive measures such as data selection, outlier detection, in-house development, and in general, privacy, security, and ethical considerations, which encompass the development stage of the AI systems, the presented technical criteria in *Table.6*, to some extent, can be offered to the willing AI developers. Furthermore, based on experts' opinions and detailed evaluation of their compliance and commitment to the determined level of security, their product can be certified. Hopefully, this viewpoint will bring the trust of the public for the AI system and gradually encourage more developers to improve the quality of their product. This report recommends that the result of this criterion or similar considerations is a measure that is scalable and is not solely based on the mutual trust between the policymakers and the developers, which will allow us to tackle cyber-threat as well as transparency issues related to the AI systems.

2.6 AI inspector

One of the major problems in AI is the lack of specialized inspectors that can certify the safety and transparency of an AI system. Currently, many AI institutes such as Mila are offering courses that generally familiarize students with problems such as bias and the interpretability of AI systems. However, the main problem here is that in such institutes, there are very few courses that are solely dedicated to recognizing ethical or technical biases in AI, and even fewer that are application centric.

Since AI applications are embedded in almost every domain, we need specialized investigators that are capable of certificating the safety of an AI system with a view that is a combination of data mining and expert knowledge about the potential ethical/ technical biases. In the last few years, training multidisciplinary students in a centralized AI institute is one of the strategies that Quebec and the AI community are promoting [108].

In AI institutes such as Mila, a student will receive an education that helps him to learn about the potential biases of the AI systems in a data mining perspective or some general concepts on the idea of ethical problems in the AI systems. Basically, they learn why an AI system should be interpretable, but not specifically what are the biases in the different AI platform combined with the internet of thing. The training that these students are receiving is not directly an answer to the biases with consideration of a specific application. Therefore, the view is broad rather than focused, while the industry may need a moderate combination of both.

The output of these institutes will be students that can design an AI system for a certain AI platform but not necessarily how to inspect its biases because they are biased to design by considering the potential biases and are not trained for the sole purpose of criticism.

Credible organizations such as Gartner also weight in importance of AI inspectors: “Promote people skills. Fill or hire people in key AI roles related to AI ethics, governance, and policy. Look for privacy/brand remediation and AI behavior forensic specialists who can explain models and perform investigations when AI fails to reduce risk [109].”

Gartner also predicts: "By 2023, over 75% of large organizations will hire AI behavior forensic, privacy and customer trust specialists to reduce brand and reputation risk [109]."

Achieving reliable AI has been one of the goals of the Quebec government and this regard there has been major investment in developing the AI ecosystem of Quebec as well as Canada [108]. In this regard, Quebec’s AI established a steering committee through the University of Montreal in order to promote

mathematical literacy and responsible AI among AI pupils and deliver specialized AI trainees to the industry [1]. However, responsible AI has received less attention since its formula may seem more dubious.

Companies such as Microsoft, Google, Samsung, IBM, and Quantum black are establishing partnerships or integrating themselves within universities to promote a future for responsible AI and the advancement of AI-based technologies.

In this regard, HumanIA in UQAM is a research-based laboratory that is focused on ethical and legal issues related to AI systems. While considering multidisciplinary issues in their defined objectives, not enough attention has been given to finding biases with regards to domain-centric AI. Considering quality assurance of the AI systems by the AI developers is already being practiced in AI laboratories and institutes such as MIT-IBM Watson AI lab, Mila institute, Gerad institute, and HumanIA. However, the main focus in these institutes is on the training responsible developers, not an AI ethical/technical inspector.

Training AI inspectors will make us ready for the upcoming changes in AI technology, prepare us for stable governance of these systems, and accelerates the process of integrating AI systems in more aspects of the technology.

These potential inspectors should possess specific characteristics that is mentioned in the following:

- A trained inspector must have strong mathematical literacy to be able to recognize the deficiencies and common biases specific to an AI algorithm.

- These trainees must be of a multi-disciplinary nature that is specialized in a few essential ground models that recognizes different types of biases of AI models. Should be able to perform adversarial test, robustness test, and security test by purely mathematical and data mining knowledge [11, 100].

- A trained person should be able to introduce adversarial input that can result in misclassification by the AI systems [100]. They should be able to introduce these inputs through the available algorithm by purely data-mining knowledge, as well as experience-based input generated by the pre-defined simulators, or based on their knowledge about the application.

- They should be able to recognize what are the elements that are missing from the input/output of an AI system that can cause a legal, discriminatorily, or ethical gap [110].

- They should also be trained specifically with regards to a certain application to be able to challenge/question the types of defined input/output, and demand adding the neglected input/output by the developer.

- A trained inspector should be aware of the IoT connected privacy, transparency, and security problems concerning the specific application of the AI system. Since every AI system can include the different realm of science such as biology, engineering, and medicine, we suggest the inspectors can be more valuable if they are domain-specific.

Furthermore, inspectors can be categorized into two groups:

1. First group can become part of the AI company, help with identifying data breaches, unintended use of the AI system, or uncover undesirable bias in the system [109].

2. The Second group are certifiers that are not affiliated with the investigated AI company. They can perform model behavior forensic [109] or identify ethical problems associated with the AI system.

The tools that investigators utilize can be categorized into pre-built machine learning investigators [109] and expert-defined analysis.

For example, an AI system that does not consider a certain input/ output in a specific application can be vulnerable to potential lawsuits or malfunction. Here, an inspector with enough knowledge of the application identifies and expresses the problem to the developers or the relevant authorities.

2.7 Who is Accountable?

Here we argue that issues with accountability may be resolved through what is proposed as ethical by design AI in the *Fig.*2. In section 2, a set of ethical constraints regarding the explainable AI are provided. These constraints can be handed to the AI developers, which they can use as their reference for dealing with their clients, how they pre-process, or how they select a representative and non-discriminative set of data. In section 3, transparency of methodology is argued as the second important factor in achieving the ethical by design AI. These transparencies can be ensured to some degree by understanding the nature of attacks, employing an explainable technique and adversarial AI (as an observer), utilizing defensive methodologies such as adversarial training for improving the AI system, and carrying out security certification tests on the AI systems.

However, one cannot expect that an AI developer possesses all of this knowledge. Here we argue that these constraints and requirements must be

worked through with the willing developers as part of a certification process that allows them to understand the standards through the guidelines provided to them. These certifications will be given to them after carefully testing their respective AI system by AI inspectors with the best tools available.

This will determine what level of safety they bring/ what level of threat they pose to the public. Therefore first, it is the clients who decide whether they are willing to use a tool with a certain level of standards or not. If an AI developer refuses to provide any level of standard for its product, they also have to face the negative consequences, in a situation where their system has created a threat/ harm toward the citizens. And lastly, those who follow the standards will be protected by the standard providers, and here the policymakers will be accountable and must work through to provide modifications for the constraints that will lead to an improved, robust, and reliable AI.

2.8 AI, Contestability, and Legal Argument

One of the main problems regarding AI technologies is the lack of preparation in the legal systems to deal with the AI-related legal arguments. There have been many cases where an AI system breached the privacy of the people. These cases resulted in the argument to be brought upon the court of law, where the AI system affected a career [16], privacy [111], or even a person's freedom [15].

People, organizations, or government harmed from a potential AI sensory device or an AI algorithm used by media platforms such as Facebook or Amazon have to utilize the legal route that is not necessarily prepared to evaluate the AIrelated cases. The missing aspects of legal route can be considered in terms of lack of evidence areas such as training, data selection, and decision making of AI systems. Lack of evidence and lack of investigators to address the concern over these criteria are amongst the main problems that we address in this note.

Contestability is introduced as one of the core principles of AI systems in numerous AI ethics framework studies [112]. It is recommended that we need a specialized legal platform dedicated to AI systems, such that people can legally challenge the abuse or harm caused by these technologies [8]. What can we do to bring more transparency to AI systems? How can we update the legal systems to be able to deal with issues that arise after the controversial decisions that are made by an AI system? What is considered as evidence?

In [113], some of the characteristics of an AI legal framework is presented as the following :

- Presenting proof of the malfunction of the AI systems.
- Contesting and correcting an error.

- Human understandable explanation.
- How much the decision made by the AI system affects the person who is contesting that decision.
- Explaining the decisions should extend to private AI decision-makers.

Here we want to elaborate the ways that evidence can be presented and offered for the legal procedures:

- An application that can have a non-negligible effect on an individual, civilians, organization, or government should be able to provide a document that includes every controversial decision plus an added explanation for that decision.
- One way to generate this explanation is through the explainable AI methodologies.
- Explainable AI methodologies are a class of tools for which a supervisory algorithm oversees the decisions of an AI system. It provides explanatory reasoning regarding the decisions made through the inner layers of an AI algorithm (known as black-box) through methods such as decision trees [29].

But how the transparency based technology will effect the future of AI Justice system and legal accountability:

- We should be able to answer the problems such as unintended use of AI system for criminal activity. Demand accountability from AI systems that knowingly shares sensitive information with third parties.
- We should also be able to demand accountability from AI technologies that can potentially cause a high level of health and safety risk for citizens, such as AI in medical applications [114] or AI in autonomous systems.
- AI-based emerging technologies will bring challenging lawsuits and may impact non-AI related trials in courts.
- We need a system with specific attention regarding AI-based technologies that do not take space, interrupts, jeopardize, and collide with the legal process of the other existing lawsuits.
- Education: we may need Juris doctorates, that besides the law are also familiarized with the bias in AI systems.

- We also recommend establishing new laws that address the violation of ethics on the emerging technologies.
- AI developers should be able to provide evidence for their argument regarding the logic of their system in a humanly understandable form. However, analyzing this evidence needs specialized trained investigators.
- Investigators of the AI systems not only can follow the presented argument by the developers but also can identify what type of information could have been included in the reasoning that was neglected by them.
- We need investigators who can analyze the evidence and allegation brought upon the court, identify the deficiencies, and present them to the person appointed that would pass a judgment on the case.

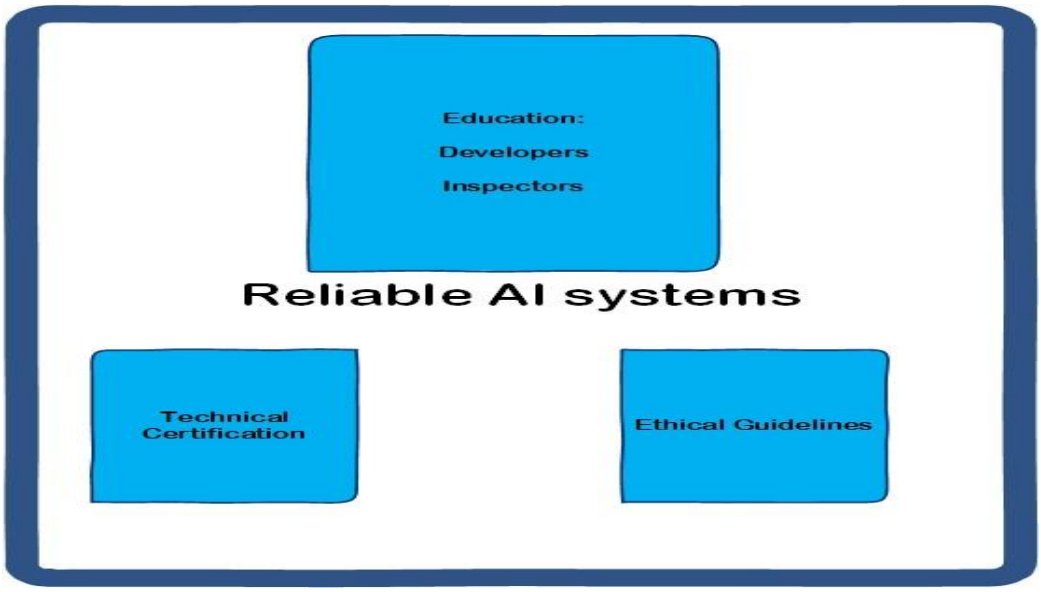

Figure 3: Necessary Elements in Regulation of AI systems

3 Ethical AI: Improvement of IoT and Emerging Technologies

3.1 Edge Cloud

In the previous sections, the characteristic of an ethical AI is investigated. However, an aspect of this issue is to approach it with more emphasis on cloud computing. In this regard, with the growing importance of processing time, lag,

cybersecurity, and the depending on the importance of the tasks, soon we will witness the prevalence of the edge-cloud technology [115].

Currently, a few companies such as Google and Amazon manage the major data centers for today's information distribution, which means due to the centralized nature of the information distribution, the data travels through many routers. With the exponential growth of AI systems, and push for the prevalence of smart technologies such as autonomous cars [116], these companies may not be able to efficiently fulfill their tasks, and issues such as lag in information transfer may impede technologies that need high-speed delivery of information. Therefore, industry and academia are pushing toward edge cloud systems, which are distributed data centers that make the processing and computing closer to the smart devices, which also include mobile processing nodes such as autonomous cars as fog nodes [116]. Therefore, every smart technology that uses an AI-based system will also be affected by this change.

Furthermore, edge cloud is a combination of technologies such as NFV (Network Function Virtualization), SDN (Software Defined Networking), and IoT (Internet of Things), which means security issues for these technologies still exists in the edge cloud and can affect privacy, security, fairness, and ethical challenges that we face for the AI systems [117]. Besides, due to more experience and a better security system, a centralized cloud owned by google might not suffer from the same issues that a local edge computing node will suffer. Also, edge clouds are more prone to physical attacks [117].

3.2 Policy issues regarding edge computing and how can we resolve them?

Edge computing that is also referred to as fog computing is regarded as a complement for the current cloud computing. The prevalence of distributed data centers means processing will be closer to the end-user. Here, we aim to investigate the risk and challenges related to this development and study its effect on values such as privacy, trust with regards to AI systems? We also want to investigate how centralized information distribution can be a threat to the AI systems, and how can we elevate this threat?

Here, the issue is that when we look at the AI policies merged with IoT, most of the current work is considering the centralized data centers, while a futuristic vision should consider the role of small businesses that will be potentially in charge of the distributed data centers [117]. To address privacy, security, and accountability in AI with edge computing, we might invest in policies that are suitable for the new era [117].

Another aspect of edge computing is authentication-based performance [117]. For example, a user allows an AI system to process its information, what fog nodes this AI system chooses to process the users' info? Should the user be informed about this? To what extent? What is the security level of these fog nodes?

For resolving issues with trust, one possibility is reputation-based trust models [117, 118]. When a user is utilizing an AI system, he or she must be informed of what type of cloud servers are used for processing the information, and the level of security standards of the distributed data centers should be transparent to the public. For example, an autonomous car company that is utilizing an AI technology should be encouraged to obtain at least a grade-based security certification from trustable industry-standard providers, which ensures the customers that their utilized data servers are robust against attacks such as a jamming attack, the man in the middle attack, and insider attacks [115, 117].

There are also issues related to fairness, privacy, and ethics. For instance, we should make sure that the server allocation of a company in terms of price, security, and data gathering is fair with regards to different regions. As an example, in [116], fairness feature for the fog based vehicular crowdsensing is mentioned, which means we assure the customer receives security assurance, fair prices, and privacy preservation, in a situation where their vehicle is being utilized as fog nodes.

We have to also pay attention to issues regarding transparency: information that is stored by these data servers should be transparent to the user. In this regard, issues such as location privacy, and data privacy are of great importance [117].

In general, neither fog nodes nor service providers should be fully trusted [118]. Therefore, we have to look into existing practices such as reputation-based models in trust models to see how small business can take part in the future of the cloud computing and how can we do this task efficiently.

In the following, a more detailed aspect of privacy, security, and fairness issues regarding edge computing is provided. Therefore, first we categorize the issues that this network upgrade presents, then we move on to explain how these issues will extend to AI systems.

3.2.1 Location Privacy

Based on [117], in edge computing, we must ensure the location privacy of the fog clients. Meaning a fog client such as an autonomous car transfers its tasks to the nearest fog nodes, and if this client uses multiple fog nodes for processing its information, it will lead to the disclosure of its estimated location and trajectory

[117]. This paper suggests the best way to resolve this issue is the "identity obfuscation", which means the fog node will not be able to identify the nearby client, as each time it will be assigned with a different fake ID [119].

Furthermore, choosing a fog node depends on criteria such as [117]:

(i) Reputation

(ii) Latency

(iii) Load Balance

(iv) Criticality of the task

3.2.2 Usage Privacy

In [117], they argue that usage privacy is another important issue directed at edge clouds. A fog node can collect the statics of the end-user, and this data might reveal their identity, the schedule of the designated task, etc. This issue has also been a focus in applications such as smart grids, where a smart meter might reveal when a client is/isn't home. In this paper, they suggest using dummy tasks for an offline client in order to deceive the adversaries, which can also be an issue of increased payment and waste of resources and energy [117].

3.2.3 Data Privacy

Privacy-preserving at the local nodes without performing decryption is one of the suggested methods in [117] to ensure the privacy and protection of the users' data.

3.2.4 Access Control

Reliable access control to ensure the privacy and security of the user with usually cryptography methods.

3.2.5 Intrusion Detection

Just like the centralized clouds already in use by companies such as Google and Amazon, fog nodes should employ robust intrusion detection systems to minimize the chances of successful cyber-attacks, such as DoS attacks, port scanning, insider attacks, and distributed attacks [117].

3.2.6 Secure Data Computation

How to preserve the computation performed at the fog nodes?

- Verifiable computing: In [117], they mention that external parallel computing performed at untrusted nodes must be verifiable [120]. Their solution is that if they utilized a function and formula at the node destination, this node must provide a log that explains why the computation is right.

  However, this put an extra diagnostic cost on the main fog node. What happens when the performed computation is a complex black box model?

  Wouldn't the verification have more computational, security, and privacy cost itself than performing the computation at the main edge node?

- Data Search: sensitive data belong to the client must be protected, issues such as what is defined as private data might also apply here. That is why controlled searching, and effective data utilization are suggested in studies such as [121].

3.2.7 Network Security

Due to the use of wireless communication for fog nodes, network security is an important aspect of the operation, especially concerning small businesses. Adversaries can use attacks such as a jamming attack, and man in the middle attack to disrupt, cause financial, mental, image related, or physical damage to the clients working with fog nodes.

There are many methodologies that are already developed and being developed such that they can help companies with network monitoring. Methods of traffic isolation and prioritization are one of the solutions for network monitoring, that are mentioned in [117, 122].

3.2.8 Trust and Authentication

Similar to prominent companies such as Google that sell the users' data to third parties, fog nodes can belong to different companies that possess different agendas. Therefore, clients need to be able to trust the fog nodes that he/ she authorizes to access their data.

Another issue is the fog node chosen by the client might use distributed processing with non-secure nodes, to perform a specific job. Therefore, what existing methodologies can we use, and what are alternative solutions that exist to resolve this problem [117, 123]?

3.2.9 Data Storage

A client must be aware of data that is stored about them at fog nodes. They should be able to inquire whether their data is being sold to third parties or not.

[117], mentions combined homomorphic and searchable encryption as a way to protect the users' data in the fog nodes.

3.3 How Fog computing affects the AI systems?

Fog computing is the enabler of cloud and edge computing (A term sometimes used interchangeably for fog computing [124]). It is the distributed processing ability provided by local data centers close to the edge device, that will form the next generation distributed local cloud computation services.

Transparency, ethics, and fairness issues of AI systems with regards to IoT are entangled with thousands of cloud servers that are operated by big companies such as Google and Amazon. However, with the prevalence of Edge-Cloud computing, the number of cloud servers will be increased to millions. Here, we try to approach how would these issues will translate to the new era of computation technology.

Ethics, privacy, security, and fairness are said to be the characteristics of an ethical by design AI [7]. Ethical AI is an AI system that is transparent, secure, and self-explanatory, such that it can aid with the legal problems that result from a decision made by an AI system. However, how can we monitor and promote ethical and transparent AI in the light of new challenges such as fog computing?

Where cloud servers and IoT multiply and becomes distributed, the potential for a more powerful AI system increases. Therefore, with capabilities offered by fog computing, training capacities, speed, and processing ability of the AI systems also increases.

A recent demonstration of the integration of 5G and edge cloud showed how machine learning models on autonomous drones, can be used to recognize humans. This matter strives toward a future where machines and humans can work safely together.

In [125], a few examples of the integration of AI and edge computing are mentioned that are presented in the following:

- Retail: Fog computing can further help personalized recommender systems. It is suggested to use video image and sensor data to further improve the recommendations that a person receives on his mobile phone about the products of his local neighborhood shop.

- Healthcare: Using edge-cloud for performing basic triage while communicating with the hospital through ambulance, already practiced in countries like China.

- Mining and resources: Fog computing will result in faster data analytics that can further improve our experience with autonomous vehicles, elevates safety, and improves applications that are able to detect and notify about the dangerous incidents.

These examples that are provided by an Australian company named Optus Business, they also suggest that edge-cloud technology may be available until 2021 [125]. Therefore, this new technology with unexplored potential should be regulated beforehand to minimize undesirables' outcomes.

Integration of 5G and edge cloud can be valuable in a situation where a stakeholder must make immediate decisions based on the data being processed [126]. In [126], a solution to respond to health crises of the COVID-19 pandemic is provided. This study provides a framework based on the integration of 5G and edge computing, which enables mass surveillance to monitor social distancing, control mask-wearing, and reading body temperature of people.

Such experiments can cause a problem in terms of ethics, invasion of privacy, and also can be used by adversaries to target individuals for their malicious purposes. How can we become proactive, to avoid any privacy and security issues?

The integration of AI and edge-cloud processing will massively enhance facial recognition technology and may be able to increase public health security [126, 127]. It may also result in an increased invasion of privacy by these systems. It enhances deep learning in such a way that it has access to faster and more broad data for its training. Therefore, the AI system itself becomes more enhanced, while the problem with the explainability of these AI systems still exist [9]. In facial recognition case, [128] describes three issues as the main problem facing the edge cloud systems:

- Data integrity: preventing data tampering by malicious adversaries.

- Maintaining confidentiality of stored and transmitted data: for this purpose, this study suggests the use of encryption techniques. However, having encrypted data is not reliable insurance for protecting the confidentiality of data. Besides, encryptions are ranged from weak to military-grade [129]. Therefore, this is a matter of cost versus privacy perseverance.

- Identity forgery: Monitoring and protecting against identity forgery is recommended. In identity forgery, attackers disguise as the authorized users. Therefore, user verification should a vital aspect of any edge-cloud network.

Problems with adversaries still exist: where they may be able to access the wireless communication between fog node and edge device. For example, an adversary accessing an edge-based facial recognition system would be a great threat to the privacy and security of the citizens, since it may be able to track individuals. There should be regulation that limits and monitors authorization of the utilization of AI technologies by the companies. For example, the combined edge cloud and surveillance system may be able to provide the ability for a company to constantly monitor people in a neighborhood. Therefore, the purpose of access to an AI system to fog nodes should be transparent.

Access to edge-based servers with certain AI systems, must be limited to trusted companies that go through an evaluation process. Otherwise, with the installation of some security cameras and a facial recognition system, any company may be able to establish its own surveillance system.

An evaluation process should consist of security checks, purpose identification, relevancy, security, and necessity:

- Security checks: if a company wants to access the edge-cloud, for tasks related to applications such as autonomous vehicles, UAVs, or facial recognition systems, their owner should be verifiable. Meaning, we want to make sure to avoid foreign adversaries using Canadian technology and use it against itself.

- Purpose: unless it is a matter of national security or public health security, AI systems such as mass surveillance may not be justifiable.

- Relevancy: facial recognition system should not be the number one primary technology of a retail company.

- Necessity: If it is not a necessary requirement for a company's AI system to access edge-cloud servers then the security risk might overtake usefulness.

- Security: evaluating AI systems capability to handle adversaries such as man in the middle attack, insider attack, and distributed attacks.

Collaborating companies that are involved with a user that utilizes an AI such facial recognition system using the edge-cloud technology must also be transparent.

3.4 Fairness and Privacy Dilemma in Personalized AI Sensory Systems with regards to 5G and Edge computing

AI sensory systems are a type of AI that utilizes a image, sound, or other sensors to collect information from an environment and based on that will result in a decision output.

Personalized AI sensory systems such as face recognition or voice recognition systems, when merged with IoT, can be used to form a powerful computation device, and with the development of technologies such as edge cloud, 5G, or combination of 5G and edge-cloud, networks can receive a boost that will help the better performance of such systems. Through this development, there may be an added efficiency in AI sensory systems since they can process more data in the same or less amount of time.

AI sensory systems, especially those that perform as face recognition or/and voice recognition have raised great privacy, fairness, and security concerns [130]. Recently countries such as France are using face recognition systems to control masks wearing in the rise of COVID-19 [131]. However, they claim that they are not storing private information [131].

Face recognition or voice recognition systems, with the help of machine learning, can also be trained to learn about the objects, humans, and the real-world. These technologies can be utilized to assist people with disabilities. People with disabilities can use these devices as it improves their livelihood through recognizing voices such as their relatives, friends, or employers.

However, they introduce individual and societal privacy and fairness concerns that need to be addressed [130]. These systems while beneficial for people with disabilities, for training and processing, need to collect data through cameras, microphones, and etc. Here, the level of transparency, security, fairness, and explainability will determine the level of trust and confidence that can answer to the concerns regarding the prevalence of such devices.

People with disabilities have the right to equality, which means if possible, they should at least have a near experience in life, similar to other people. It may seem only ethical for them to be provided with a certain level of amenity through possible technologies such as AI sensory devices. It is not fair to a person who is deaf or blind to be denied the benefit of AI sensory devices. Through these devices, they will able to regain simple advantages such as easily understanding knocks on doors or being informed when a familiar face is close to them [130].

For example, the concept of a smartwatch for the visually impaired person is discussed in [132], where these individuals can attend a meeting that requires silence, and at the same time, they would be able to recognize the speaker and the people present in that meeting without disrupting the meeting. In [132], their prototype does not use the internet for boosting search results.

However, these amenities are almost useless for a human that is not blind or deaf. Besides, non-disabled people might feel uncomfortable knowing a camera is watching them or a device is recording and processing their voice [132]. Therefore, there exist important questions about individual and societal responsibilities[130]. For example, a user, while benefiting from a face recognition system, might be wary of inadvertently sharing personalized sensitive information.

How can individual privacy of those in close contact with such systems be preserved? If they are employed for personal use, how advanced these AI systems should be? In a facial/voice recognition case, do we want them to recognize every face/voice of a celebrity, online friends, or strangers [130]? What information should the result of their decision making provide, and what are the input data in such devices? As a Face recognition system may also read sensitive data such as credit cards. What if a person with a disability has surveillance and malicious motivation when using AI sensory systems[132]? How can we avoid biases, especially toward already marginalized groups [132, 133]?

3.5 Possible Solution to Privacy and Fairness Issues Regarding Personalized AI Sensory Systems

Based on the what we have approached so far about data privacy, explainable AI, and IoT privacy issues regarding the edge system in combination with AI systems, we recommend the following practices for dealing with privacy and fairness for Personalized AI sensory systems, although some of the recommendation might be extendable to Non-Personalized AI:

- Fairness in data training, we should be able to determine who decides the inputs and how are they chosen, and what label would the data will be assigned [130].

- There are already existing all-party consent laws in place, which limit the development of such devices [134]. These laws that are already established in countries such as the U.S or Canada need to be carefully modified, and studying them will help us to determine the appropriate measures in terms of preserving privacy for the utilization of such devices.

- We should also monitor the companies that manufacture these devices. Adding security certification based on their provided level of transparency is the first thing that we should implement when dealing with such systems. In this way, we can offer a level of trust to the users.
- Unauthorized access of these systems to the cloud database must be forbidden.
- We should encourage built-in AI sensory systems for personal usage, where access to the internet and global search engines are limited. If these systems utilize the internet and high-speed technologies such as 5G, then we recommend that their users should be identifiable. For example, a visually impaired person that uses the cloud database should have an identifiable code.
- We should identify and certify what type of information the database for these technologies will store. Information such as name, age, height, and credit card. Long term memory lessness of these devices should be one of the aspects of their product.
- We should examine through practices such as technical AI certification that is mentioned in the previous sections, that the information that such systems accept as input, does not result in a discriminatory output.
- These systems will constantly upload information as they are trying to be assistive to their user. There should be a built-in mechanism that prevents uploading sensitive information such as credit card details since the excess info can become accessible to every collaborating third party companies with the cloud.
- We should also invest in cyber-security measures that deal with adversaries that impersonate a user.
- We should also consider security measures that protect users' information.
- We should also consider a potential user might be an adversary.

3.6 Categorizing AI Product Based on Their risk

For accomplishing a coordinated investment to address the ethical, privacy, and security aspects of these technologies, there exists a need for

studying the policies for the AI-based product based on the categories that can represent their threat.

Here we discuss that the categorization depends on a variety of parameters such as user's age, type of application, and the purpose of the AI application. We also explain how context matters and higher intelligence level in AI systems does not necessarily mean a greater AI threat.

In [135], it has been stated that AI technologies can be categorized through three dimensions, such as multi-functionality, intelligence, and user interaction. Each of these dimensions can be subjected to ethical, security, transparency, and privacy concerns. The purpose of this categorization is to help direct the efforts for tackling the AI issues.

More interaction means more threats since the AI system needs a higher level of features to be able to improve the user interaction [135]. For example, in infotainment applications of the smart vehicles recommendation system, besides the search history and user's preference, the location of the vehicle may be used as well [136].

Multi-functionality also poses a great threat, since it means AI device is collecting more sensory information. For example, smartphones or smartwatches, depending on the type of information they collect, such as voice, image, search history, and location [135], can be subjected to ethical, privacy, and security issues. In [135], AI intelligence is also introduced as another dimension in which a more intelligent AI system is presumed to be more threatening [135].

However, a more intelligent AI system does not always mean more threats. Two AI products that have the same level of access to sensory devices such as cameras and microphones with internet accessibility can present the same level of security threats. The security threats depend on how much preventive and defensive mechanism an AI system offers.

We can also define a more intelligent AI as a system that also considers the security measures. Here, we want to introduce a different way of categorization, and we recommend that targeted regulation based on AI systems' special features, their user, or special application may lead to a better path for coordinated investment and tackling AI-related issues. We need to make the path for policies clearer rather than obscure.

Considering the following type of regulations will help us to categorize the AI systems based on their ethical, security, or privacy concerns:

- Regulation based on the type of information that the AI system uses. For example, evaluating the type of sensory devices that the AI system utilizes.
- Organizational specific policy. For example, we have to make sure private organizations do not utilize AI systems for employee behavior monitoring applications [137].
- We need user-specific policies, where the AI system is regulated such that a more vulnerable user will be offered more protection. For a regular citizen, location information with regards to infotainment applications does not create a high level of security concerns. However, for high-profile officials, we have to be warier of adversaries.
- What age groups does the AI product targets? For example, the use of facial recognition systems in AI products designed for children should be of special concern.
- How much the AI system is internet dependent? What type of information is transferred and processed through the internet?
- Regulating the AI systems based on their purpose. For the AI system in the medical domain, transparency and ethical issues are the concerns. AI systems used in the vehicular ad-hoc network (VANET) can be a danger to the safety of citizens [136].
- Regulation based on the reachability of the AI systems. We need policy specific for AI systems that can be manipulated and used as a national threat. For example, recommender systems in social media applications.

3.7 Smart Cities

Technological advancement aims to synergize autonomous systems such as courier robots and autonomous vehicles with the latest developments of AI and IoT. For example, courier robots are a kind of device that companies like Amazon are considering for wide-scale use for transporting merchandise through unmanned aerial vehicles. These technological advancements will form a structure that is known as a smart city, where we become closer to automation in many areas such as agriculture, disaster management, and transportation. In this section, we present a recommendation for establishing policies in regards to security aspects of critical facilities and components within a smart city.

Smart cities will be utilizing smart solutions in several areas of agriculture, education, transportation, maintenance, education, governance, smart industries, smart healthcare, smart energies, and smart policies [138]. The different components of smart cities can be considered as cloud/edge computing, IoT (internet of things), machine learning and Cyber-Physical systems, security protocols, wireless sensor networks, 5G, ICT (Information and Communication Technology), and geospatial technologies [138].

Here we aim to ask policy questions on the interconnection between different sections of smart cities in order to develop a unified framework for protecting smart cities against adversaries. A few considerations for developing such framework are established and reviewed through the following studies:

- In [139], technical challenges are categorized as security threats, interoperability, lack of supporting infrastructure, unstructured data management, and absence of universal standards.
- In terms of navigation and threat monitoring, smart solutions are suggested, such as smart air monitoring through UAVs and satellite data [139].
- Current international standards for smart cities are under ISO, IEC (International Electrotechnical Commission), and ITU (International Telecommunication Union) [138].

Furthermore, we present our recommendation for policy questions regarding for preserving security, privacy, and transparency in smart cities:

- Developing security models for smart city management that consider the system level and component level defense including sensors, actuators, networking, and communication. An example of these security models is presented in [140].
- Focusing on privacy models that consider the interaction between components:
  - How does edge computing affect the policies regarding UAVs?
  - What type of information should be picked up, and transferred by the sensor network of the AIoT device and its shared data centers.
  - What type of info should be memorized in the destined data centers.
  - What type of AIoT devices can harm a critical CPS infrastructure?

  - What type of AIoT devices are a threat against officials?
    - Strict policies for swarm-bots.
    - Strict policies against surveillance devices.
    - Strict policies regarding officials on the utilization of AI application that are originated from countries that have not addressed the accountability issue of the questioned AI system.

- What type of collaboration between AIoT devices can lead to location identification and strategic information extraction?
- Monitoring systems that are specifically designed toward supervising air traffic and ground traffic caused by autonomous systems.
- Finding ways to monitor unregistered sensory devices, such as working toward spoofing detection [141].

4 A closer look at AIoT based Emerging Technologies

In this section, the aim is to introduce a different variety of emerging technologies, and highlight possible privacy, fairness, transparency, and ethical issues with respect to that technology, and provide recommendations in order to be proactive and ready, so that it can result in steady governance of emerging AI-based technologies.

4.1 Security and Privacy Issues in Vehicular Cloud Computing

Vehicular Ad hoc Network (VANET) is one of the ideas that is developing with the advancement of IoT. VANET creates an Internet of Vehicles (IoV) such that the vehicles are equipped with sensory devices that provide computation, storage, or networking [142]. VANET based systems will have a variety of applications such as entertainment, navigation, or distaste control. The effectiveness of these systems depends on low latency, a reliable communication network, and real-time data-processing [143]. With the propagation of 5G technology and becoming closer to the emergence of edge cloud computing services, the automotive industry will be promoting and implementing intelligent and novel applications of VANET. We will briefly mention some of the benefits of this technology, and also the possible challenges in terms of security and privacy.

Vehicular cloud network connections can be between the vehicles or from the vehicles to the designated cloud computing infrastructures (for example, a local fog computing server) [143]. The vehicular cloud network will reduce the expenses of IoT computational resources for the vehicle manufacturer.

VANET will provide a distributed infrastructure that will be utilized for data storage, traffic management, surveillance, or infotainment purposes [143]. It has application in traffic management, where an unexpected incident due to an earthquake, car accident, or repairs can be reported through a network of cars [136, 143]. It has applications in surveillance, where sensory devices can be installed on vehicles with image recognition capabilities that can be used for a variety of purposes, such as reporting suspicious activity, location detection, and emergency broadcasts [136]. It has applications in infotainment, where the client's data will be utilized for distributing information about the road, weather, or entertainment purposes for the driver [136, 144].

Safety regulation specific to adversarial based attacks, privacy concerns regarding location, shared information between the vehicles and networks, and issues with usage of image recognition devices are of the main concerns for this technology.

In order to approach the policies, first, we have to categorize the special features that are relevant to these type of systems:

- They are capable of using image recognition systems.
- They can be connected to each other as well as a neighboring local network.
- They are also connected to the global network in a distributed manner.
- They are prone to adversary attack with intentions to cause injury, stole private information, or distributing false information.
- They are deployed on the road, and they are in contact with civilians.
- They can lead to location identification.
- They constantly share and receive information to be able to perform their designated tasks.
- Their application defines the data that they need to process, so in each case, they will need a specific type of sensory information.

Here, with consideration of the above elements, we present some of the recommendations that will help us to regulate this technology:

- We have to tackle privacy, safety, and security problems with respect to applications: meaning for each application, such as infotainment, disaster control, or surveillance, different policies have to be considered.

- We should find an answer to the privacy challenges of VANET systems with respect to the IoT. We should categorize the policies to the vehicle to vehicle and vehicle to cloud policies.

- For example, besides the vehicle to cloud protocols, vehicle to vehicle communication protocol must also follow high standards to prevent adversarial access.

- Large-scale utilization of image processing technology should be closely monitored. For example, in surveillance and disaster control, only trusted or government-supervised companies should be allowed to perform.

- An AI surveillance system should not be used as a judgment tool on the intent of a crime. This type of technology can cause damage to the trust of the society toward the government. Therefore, it must be ensured that these systems will gather and store information relevant to their intended use.

- We must provide a clear definition of sensitive data. For example, VANET technology can help in situations such as finding lost children. However, gathering image data from children can be considered problematic.

- We should answer the safety challenges with respect to the IoT and transparency of these type of technologies:

- Are these systems are tested and robust against adversarial attacks? For example, can an adversary take control of the autonomous?

- Companies must be transparent in terms of the type of data that they store. There should be a mechanism in place that would enforce the removal of sensitive data. Edge servers also should be prohibited to share or sell sensitive data to third parties. However, the servers may be allowed to gather and buy non-sensitive necessary data from third parties to boost the performance of such a system.

- The above means the necessary information may be bought but should not be sold.

5 Impact of AI and emerging technologies in future warefare and policies

5.1 IoT and technological impacts

In this section, we want to investigate the technological impacts of IoT devices in terms of military application. How can we make sure we can trust the communicated information, neutralize the threats, and improve transparency? For this purpose, we analyze two aspects of this problem that is network solution and the AI-based Application solution.

5.2 Zero-trust network policy

First, we want to investigate that how can we achieve reliable and secure connection of AIoT devices in military applications. For example, cognitive AI is known to be the next-generation approach to AI systems that can constantly adjust and adapt to situational events. Adapting to situational events is a prominent feature of military applications and can be used for both domains of AI cyber defense and offense [145]. The adversary may use advanced technologies such as cognitive AI for manipulating, disrupting, or targeting ally resources and AIoT devices. In this regard, we address the zero-trust architecture as a possible way for countering advanced AI-based adversaries in battlegrounds.

Based on [146], integration of advanced identity management, software-defined networking, and hybrid multi-cloud capabilities is deemed to provide the fast and reliable cyber platform needed for implementing military strategies in a zero-trust network architecture. In [146], it is also mentioned that futuristic military-based cyber platform needs novel data-science algorithms, while we must also make sure the current ones operate with maximum security. In other words, we have to adapt to the notion of "verifying and never trust."

One solution is zero trust security architecture. In a zero-trust security architecture, the users are connected directly to their respective devices [147]. However, there is a trade-off between connectivity and security when it comes to zero trust network architecture. Here, we mention some of the features of this strategy and try to recommend policies that can be helpful for improving network security in a rapidly changing environments such as battlefields.

- In [146], it is mentioned that a zero-trust cyber platform must have certain characteristic such as the following:

  - Software-Defined Networking and ICAM (identity credential, access, and management) must be ensured to be of a zero-trust nature. For example, every device in the network must be identifiable.
  - All networks must be assumed to vulnerable to manipulation.
  - Users should only have access to their respective needed resources.
  - We need Real-time detection and protection capabilities.
  - Maintain situational awareness.
  - Must be standardized and certified.
  - Ready for fast response to emerging ISR.
  - Support Multi-Cloud and edge computing.
  - Having a modern and programmable software-defined networking such that they can enforce new policies.
  - Operational agility and have flexible options network maneuver.

- Organizational theories concerned with zero-trust [148] can be extended to set policies for the zero-trust architecture in military applications.

In the following we mention a few recommendation as the necessary element and considerations for establishing a zero trust network architecture:

- A zero-trust policy should follow a hierarchy starting from the received input from top commanders to shared info between the army units and the integrity of communication between the operational AI-based military equipment. The aforementioned point is mainly necessary to avoid the risk of impersonation and manipulation of IoT-based exchange of information.

- In a zero-trust policy, no asset is trusted. Therefore, policies for the development of human-aware AI (cognitive AI) and trained professionals that are aware of their controlled AI device should be established.

- The potential for capturing assets and reverse engineering them should be considered in the design stage [148] .

- Developing new deep learning algorithms that have better observability and threat verification capability than the current ones should be considered. Current black box models may not be observable across all their entry data such that with knowledge of output, we can have an idea of the entry data to their models.

- Policies that encourage applying, converting, and integrating organizational theories concerned with zero-trust into the battlefield framework while also reinforcing and improving the already existing policies [149] .

- Risk assessment in finding how hybrid-trust in civil applications can negatively affect national security. Here, the hybrid trust means in some applications the zero-trust policies are considered, and in less critical applications such policies are relaxed to ensure IoT services are fast and less disrupted

5.2.1 Cognitive AI

As the nature of adversarial attacks is rapidly changing, there exists a need for an AI security defense mechanism that is inherently aware of its own uncertainties and can integrate with next-generation technologies such as edge computing to enhance military applications. Here, we want to take a closer look at the next generation technology for security operation centers that is referred to as cognitive AI.

Since IDSS systems are vulnerable to adversarial attack, have bias and trustability problems, and can not provide the needed flexibility of decision making in military applications, a new generation of AI security concept is deemed to enhance the operational security against threats such as phishing, malicious data tampering, and dos attacks [145]. This idea is referred to as cognitive AI, where there is a mutual awareness between AI and humans. In [145], three elements are mentioned as the foundation of such human-AI cooperation. Be mutual predictable, mutually directable, and have mutual common ground.

The cognitive AI is a next-generation technology that is discussed for improving military decision making and boosting cybersecurity defense as the traditional IDSS (intelligent decision and support system) have various shortcomings on trust and security due to the black-box nature of AI systems.

The main question is, based on an environment that the AI system is deployed, how can we achieve situational awareness [145, 150]? This an important issue, specifically, in military applications, where there can be

•

variables such as fatigue, needs, capabilities, and malicious intentions. Here, we aim to discuss cognitive AI, companies that are moving forward with this type of technology, what should be done to accomplish it, and what we can accomplish with this type of AI technology.

In terms of companies advancing this technology we can mention IBM and CISCO:

In the view of IBM QRadar Advisor with Watson cognitive AI can overcome the lack of talent and job fatigue in cybersecurity:

- It can visualize how the attack is progressing, validate the threat, and suggest what are the possible threats that can still occur.
- Possess cognitive reasoning for isolation of threat.
- Provides a priority-based investigation list.

- CISCO also provides edge and fog processing, data analysis, feedback, and computation that can be a key concern in a connected battlefield:
  - Provides solutions such as joint node networks enabling soldiers to communicate via satellite.
  - Considers the technology of mobile edge computing, which can provide the networking and interconnection between the AI devices.

Here are a few of our recommendation for achieving, importance, and accruing the cognitive AI technology:

- One of the main challenges of mobile edge computing is the security and trustability of exchanged information. Developing cognitive AI can provide this by bringing observability, explainability, awareness, and constant reconfiguration and learning for AI systems against AI threats.
- We need a combination of mobile edge computing and cognitive AI to bring situational and environmental awareness, as well as being able to establish a human-AI interconnection.

-
- For example, consider a group of scattered UAVs that are controlled by military operators and communicate with each other through military vehicles and devices that are installed/on the move in an area of operation such that together will form a mobile edge computing server. One of the cognitive AI technology's roles is to help operators to remain responsive to potential threats and assess the integrity of exchanged information between UAVs or their input/output commands.

- In terms of training the military personnel, we can collaborate with companies such as CISCO.

  We need to establish cognitive strategies that aim to resolve cognitive challenges such as massive data, a fusion of complex data, building site-specific knowledge, and maintaining multiple mental models [151].

- We need collaboration with universities in areas that we can train edge AI and cognitive AI experts so that in long term we can keep up with the technological advances that are mobile as well as secure against adversarial attacks, are explainable, and aware of their environment, and human operator.

5.3 Defensive policy impacts

Advances of AI technologies have foreseen to be transformed into an arms race, the next nuclear capability, who has it first and who becomes more advanced. While ethics and privacy may be perceived as a natural barrier to the development of certain AI cyber-offense tool, it should also push us to persevere security, privacy, and right of civilians through the AI cyber-defensive capabilities.

Here, we highlight areas that require consideration for policy development for AI based technologies that may enhance the defensive capability of Canadian armed forces.

The next generation of warfare is perceived to be enhanced with AI-based technologies. Tools that are C4ISR, whether the efficiency is high enough or not, the countries that possess AI-based warfare technologies are deemed to be

•

superior in terms of futuristic arm race, due to factors such as sensory capabilities and fast decision making [152]. Cyber-threats against nuclear systems, ISR (Intelligence, Surveillance, Reconnaissance), and robotic warfare are some of the areas that will become more advanced/threatening with the help with AI-based technologies [153].

Investing in AI-offensive warfare may seem a simpler and more convenient approach, although there can be humanitarian, privacy, and other ethical barriers that may prohibit the development of certain type of such technologies (e.g. image recognition based surveillance) The absence of robust defense policies is a significant contributor to further uncertainty in the case of AI-powered confrontation. For example, AI-based robotic warfare can provide a cheap counter against advanced technologies such as submarines or fighter jets, but how should they be stopped? [152].

Furthermore, the threat is not always international and individuals with malicious motives may use such systems for their malicious intents. Therefore, AI defense can cover a bigger and more imminent class of adversarial threats, especially

now, with easy and cheap access of adversaries to AI-empowered technologies.

In [152], they categorize AI-enhanced capabilities to:

- Digital security: against threats such as phishing, impersonation, and data poisoning.
- Physical security: swarm attack by drones for targeted assassination.
- Political security: such as surveillance, breaching, and deception.

  Political security can be viewed from two points of view, one is outsider threat, and the other in terms of authoritarian government. For example, using existing or planted surveillance systems to gather intelligence by malicious adversaries or utilizing these systems to control and oversee people's behavior by authoritarian governments.

Investing in cyber-defense tools :

- Analyzing classification errors.

  Why they happen? how can we improve them? Why different classification error can be prone to same attack?

- Automatic detection of vulnerability.

  Implementing observer-based methodologies that can detect the infiltration of AI systems.

5.3.1 Recommendations

Here are a few important areas of security against AI-based warfare tactics:

Secure cyber-space:

- Using identity authentication methods in data transfer.
- Prevalence of secure data centers for data transfers such as AZURE government and AWS government to limit the capability of outsider threat by enhancing local data centers.
- Prevalence of Robust AI-based intrusion detection system:

  - Adversarial awareness, by investing in finding and examine possible breaches into the security systems before it happens. In other words, the offense must be a way to examine the defensive measures.
  - Explainable: Investing in methodologies that provide a traceable record of the event that happened. Who? What? Where? When? Why? [154] – Investing in cognitive security analytics.

- Examining technological warfare: understanding and analyzing the counters to technological threats will make us alert for possibilities:
  - Categorizing the dual use nature of AI technologies such as UAVs, packaging, or assassination?
  - Finding effective ways for countering advanced warfare such as AIbased swarm attacks, for example, one can build a playbook based tactic and invest in AI-based methodologies that can counter them:
    * Electromagnetic based AI weapons
    * Communication Jammers
    * Hijacking the swarm with data injection attacks

5.4 International impact

The progress of countries with non-transparent policies in the advancement of AI-based technologies is one of the main concerns of governments and regulators such as the U.S and the EU. In this section we mainly focus on how Canadian should be concerned about non-transparent AI-based applications.

In each century, people's view of what is ethical is changing, therefore rules and regulations have been updated based on what we the general population perceive as ethical or moral behavior. Sometimes a catastrophic event may lead to the new regulation, and in other instances foresight and proactiveness. It is important to regulate AI-based technologies that can be potentially be used under military-civil application. Policies that demand AI developers to unconditionally share their data with the government can be a potential threat to the privacy of Canadian consumers, government officials, and in general national security.

From the EU's perspective, international collaboration with such countries can lead to better transparency and getting power from hard-liners. It suggests that collaboration will lead to familiarizing themselves with their intentions while having significant economic benefit [155].

Also, due to restrictions such as the privacy of consumers, countries with transparent policies are deemed to be at a disadvantage in terms of the development and advancement of AI-based applications in the areas such as facial recognition [156]. Countries with unlimited access to private data can eventual form a stronger facial recognition and surveillance systems than the countries that are not practising such policies.

Here the question is, how can we regulate the potentially dangerous AI-based applications with non-transparent data policies?

- For dual-use AI applications that can be potentially dangerous:
  - Promoting usage of data centers that are located within a safe-zone to the domestic AI developers. For example data centers belonging to countries that possess transparent AI policies.
  - Working toward international safe zone data centres.
  - Risk assessment to provide stricter measures and data center usage for AI-based application that pose greater threat.
- Setting a standard non-discriminative rule that demands transparency from AI developers in the case AI technologies with high risk of military-civil capabilities.
  - The rules prevents these any AI developer from using non-secure data centers.
  - Sharing data to unauthorized third parties.
  - Banning access to non-essential data for the AI application.
- Promoting Edge computing. Edge computing prevents the need for data to travel across the continent and improves data security of consumers.

6 Conclusion

In this report, first, some of the important definitions regarding the explainable AI methods are mentioned. Next, the ethical constraints for achieving an "ethical by design" AI, based on Australia's ethic framework is evaluated. Besides, a review of explainable AI methodologies is presented in Table. I, and lastly adversarial example and security standards with regards to methodology that must be considered by the developers are addressed. These standards address the issue from the perspective of attacks on the AI or explainable AI.

By considering the reviewed studies, it can be concluded that there are two aspects to design constraints for an ethical AI, which can be viewed in terms of technical AI design regulation (transparency of the methodology), and ethical standards. The developers must consider the technical aspect and ethical constraint at the same time. In terms of the technical aspect, the developer has to choose a methodology that is transparent with regards to data and must ensure that its model is not vulnerable to adversarial examples so that it does not put public security and privacy at risk. The integration of methodology standards and ethical standards will result in an AI that is ethical by design and can help the government to regulate these systems.

Beside from well-educated AI developers, to implement and regulate these aspects, our study recognized the need for the specialized AI inspectors that not only are mathematically well-educated, but they are also specialized to recognize biases that are application-specific and may go unnoticed by an AI-forensic specialist that only is educated in the data mining aspect of the problem.

In the second part of our research, we turned our attention to transparency, fairness, privacy, and security issues regarding the AI system with regards to emerging technologies such as edge-cloud computing and 5G. In this part, our purpose is to raise awareness about the upcoming flow of unique ethical challenges that we will face in the next few years due to the fast propagation of these technologies. We also provided a few recommendations that hopefully will help us deal with these challenges.

References

[1] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable ai systems for the medical domain?," *arXiv preprint arXiv:1712.09923*, 2017.

[2] E. Kocabey, M. Camurcu, F. Ofli, Y. Aytar, J. Marin, A. Torralba, and I. Weber, "Face-to-bmi: Using computer vision to infer body mass index on social media," in *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[3] M. S. Gerber, "Predicting crime using twitter and kernel density estimation," *Decision Support Systems*, vol. 61, pp. 115–125, 2014.

[4] P. van Esch, J. S. Black, and J. Ferolie, "Marketing ai recruitment: the next phase in job application and selection," *Computers in Human Behavior*, vol. 90, pp. 215–222, 2019.

[5] E. Turban, R. Sharda, and D. Delen, "Decision support and business intelligence systems (required)," *Google Scholar*, 2010.

[6] V. M. Hudson, *Artificial intelligence and international politics*. Routledge, 2019.

[7] D. Doran, S. Schulz, and T. R. Besold, "What does explainable ai really mean? a new conceptualization of perspectives," *arXiv preprint arXiv:1710.00794*, 2017.

[8] M. Zalnieriute and O. Gould-Fensom, "Artificial intelligence: Australia's ethics framework submission to the department of industry, innovation and science," *UNSW Law Research Paper*, no. 19-40, 2019.

[9] J. Zach, "Interpretability of deep neural networks," 2019.

[10] N. Dalvi, P. Domingos, S. Sanghai, D. Verma, *et al.*, "Adversarial classification," *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 99–108, 2004.

[11] G. Singh, T. Gehr, M. Mirman, M. Püschel, and M. Vechev, "Fast and effective robustness certification," *Advances in Neural Information Processing Systems*, pp. 10802–10813, 2018.

[12] D. Reinsel, J. Gantz, and J. Rydning, "The digitization of the world from edge to core," *Framingham: International Data Corporation*, 2018.

[13] report by The Canadian Press, "Bell launches 5g network in montreal, toronto, calgary, edmonton and vancouver." URL: https://www.ctvnews.ca/business/ bell-launches-5g-network-in-montreal-toronto-calgary-edmonton-and 4979943.

[14] R. van der Meulen, "Edge computing promises near real-time insights and facilitates localized actions.." https://www.gartner.com/smarterwithgartner/ what-edge-computing-means-for-infrastructure-and-operations-leade

[15] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias risk assessments in criminal sentencing," *ProPublica, May*, vol. 23, 2016.

[16] "Design, scope, cost-benefit analysis, contracts awarded and implementation associated with the better management of the social welfare system initiative." Senate Community Affairs Committee, Parliament of Australia, 2017.

[17] "Take control of your virtual identity," *European Commission*, 2019.

[18] "Westpac testing ai to monitor staff and customers."URL:urlhttps://www.afr.com/technology/westpac-testing-ai-to-monitor-staffand-customers-20171114-gzks7h.

[19] "How facial recognition technology aids police." URL: http://theconversation.com/ how-facial-recognition-technology-aids-police-107730.

[20] M. Taddeo, T. McCutcheon, and L. Floridi, "Trusting artificial intelligence in cybersecurity is a double-edged sword," *Nature Machine Intelligence*, pp. 1–4, 2019.

[21] I. Stoica, D. Song, R. A. Popa, D. Patterson, M. W. Mahoney, R. Katz, A. D. Joseph, M. Jordan, J. M. Hellerstein, J. E. Gonzalez, *et al.*, "A berkeley view of systems challenges for ai," *arXiv preprint arXiv:1712.05855*, 2017.

[22] "Developing standards for artificial intelligence: Hearing australia's voice — submission to standards australia." URL: https: //www.oaic.gov.au/engage-with-us/submissions/developing-standards-for-artificial-intelligence-hearing-australia

[23] "Regulation (eu) 2019/881 of the european parliament and of the council." URL: https://eur-lex.europa.eu/eli/reg/2019/881/oj.

[24] E. Glaessgen and D. Stargel, "The digital twin paradigm for future nasa and us air force vehicles," in *53rd AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics and materials conference 20th AIAA/ASME/AHS adaptive structures conference 14th AIAA*, p. 1818, 2012.

[25] P. Nguyen and L. Solomon, "Consumer data and the digital economy: Emerging issues in data collection, use and sharing," 2018.

[26] B. Bagdasarian, "Conversational data collection: Active data vs passive data." URL: https://blog.hubspot.com/customers/ converstional-data-collection-active-passive.

[27] T. A. Singlehurst, M. Kelley, A. Shirvaikar, C. T. O'Neill, M. May, and W. H. Pritchard, "eprivacy data protection who watches the watchers? – how regulation could alter the path of innovation," *Citi Global Perspectives & Solutions*.

[28] "The tiger mom tax: Asians are nearly twice as likely to get a higher price from princeton review."URL:https://www.propublica.org/article/asians-nearly-twice-as-likely-to-get-higher-price-from-princeton-

[29] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[30] G. Ras, M. van Gerven, and P. Haselager, “Explanation methods in deep learning: Users, values, concerns and challenges,” in *Explainable and Interpretable Models in Computer Vision and Machine Learning*, pp. 19–36, Springer, 2018.

[31] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, “Layer-wise relevance propagation: an overview,” in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 193–209, Springer, 2019.

[32] A. Rai, “Explainable ai: from black box to glass box,” *Journal of the Academy of Marketing Science*, vol. 48, no. 1, pp. 137–141, 2020.

[33] D. M. Turek, “Explainable artificial intelligence (xai).” URL: https://www.darpa.mil/program/ explainable-artificial-intelligence.

[34] D. Bau, J.-Y. Zhu, H. Strobelt, B. Zhou, J. B. Tenenbaum, W. T. Freeman, and A. Torralba, “Visualizing and understanding generative adversarial networks,” *arXiv preprint arXiv:1901.09887*, 2019.

[35] G. Rätsch, S. Sonnenburg, and C. Schäfer, “Learning interpretable svms for biological sequence classification,” in *BMC bioinformatics*, vol. 7, p. S9, Springer, 2006.

[36] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, G. P. Way, E. Ferrero, P.-M. Agapow, M. Zietz, M. M. Hoffman, *et al.*, “Opportunities and obstacles for deep learning in biology and medicine,” *Journal of The Royal Society Interface*, vol. 15, no. 141, p. 20170387, 2018.

[37] Z. Che, S. Purushotham, R. Khemani, and Y. Liu, “Interpretable deep models for icu outcome prediction,” in *AMIA Annual Symposium Proceedings*, vol. 2016, p. 371, American Medical Informatics Association, 2016.

[38] J. Kim and J. Canny, “Interpretable learning for self-driving cars by visualizing causal attention,” in *Proceedings of the IEEE international conference on computer vision*, pp. 2942–2950, 2017.

[39] B. Letham, C. Rudin, T. H. McCormick, D. Madigan, *et al.*, “Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model,” *The Annals of Applied Statistics*, vol. 9, no. 3, pp. 1350– 1371, 2015.

[40] S. García, A. Fernández, and F. Herrera, “Enhancing the effectiveness and interpretability of decision tree and rule induction classifiers with evolutionary training set selection over imbalanced problems,” *Applied Soft Computing*, vol. 9, no. 4, pp. 1304–1314, 2009.

[41] F. Wang and C. Rudin, “Falling rule lists,” in *Artificial Intelligence and Statistics*, pp. 1013–1022, 2015.

[42] G. Su, D. Wei, K. R. Varshney, and D. M. Malioutov, “Interpretable two-level boolean rule learning for classification,” *arXiv preprint arXiv:1511.07361*, 2015.

[43] D. M. Malioutov, K. R. Varshney, A. Emad, and S. Dash, “Learning interpretable classification rules with boolean compressed sensing,” in *Transparent Data Mining for Big and Small Data*, pp. 95–121, Springer, 2017.

[44] V. Schetinin, J. E. Fieldsend, D. Partridge, T. J. Coats, W. J. Krzanowski, R. M. Everson, T. C. Bailey, and A. Hernandez, “Confident interpretation of bayesian decision tree ensembles for clinical applications,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 11, no. 3, pp. 312– 319, 2007.

[45] S. Hara and K. Hayashi, “Making tree ensembles interpretable: A bayesian model selection approach,” vol. 84, pp. 77–85, 09–11 Apr 2018.

[46] H. F. Tan, G. Hooker, and M. T. Wells, “Tree space prototypes: Another look at making tree ensembles interpretable,” *arXiv preprint arXiv:1611.07115*, 2016.

[47] R. D. Gibbons, G. Hooker, M. D. Finkelman, D. J. Weiss, P. A. Pilkonis, E. Frank, T. Moore, and D. J. Kupfer, “The cad-mdd: A computerized adaptive diagnostic screening tool for depression,” *The Journal of clinical psychiatry*, vol. 74, no. 7, p. 669, 2013.

[48] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems*, pp. 4765–4774, 2017.

[49] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should i trust you?: Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, ACM, 2016.

[50] M. T. Ribeiro, S. Singh, and C. Guestrin, “Anchors: High-precision modelagnostic explanations,” in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[51] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti, “Local rule-based explanations of black box decision systems,” *arXiv preprint arXiv:1805.10820*, 2018.

[52] S. Mishra, B. L. Sturm, and S. Dixon, “Local interpretable model-agnostic explanations for music content analysis.,” pp. 537–543, 2017.

[53] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.R. MÃžller, “How to explain individual classification decisions,” *Journal of Machine Learning Research*, vol. 11, no. Jun, pp. 1803–1831, 2010.

[54] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” *arXiv preprint arXiv:1312.6034*, 2013.

[55] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *European conference on computer vision*, pp. 818–833, Springer, 2014.

[56] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” pp. 2921–2929, 2016.

[57] R. C. Fong and A. Vedaldi, “Interpretable explanations of black boxes by meaningful perturbation,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3429–3437, 2017.

[58] P. Dabkowski and Y. Gal, “Real time image saliency for black box classifiers,” in *Advances in Neural Information Processing Systems*, pp. 6967– 6976, 2017.

[59] V. Petsiuk, A. Das, and K. Saenko, “Rise: Randomized input sampling for explanation of black-box models,” *arXiv preprint arXiv:1806.07421*, 2018.

[60] G. Ras, M. van Gerven, and P. Haselager, “Explanation methods in deep learning: Users, values, concerns and challenges,” in *Explainable and Interpretable Models in Computer Vision and Machine Learning*, pp. 19–36, Springer, 2018.

[61] U. Johansson, R. König, and L. Niklasson, “The truth is in there-rule extraction from opaque models using genetic programming.,” in *FLAIRS Conference*, pp. 658–663, Miami Beach, FL, 2004.

[62] M. H. Aung, P. G. Lisboa, T. A. Etchells, A. C. Testa, B. Van Calster, S. Van Huffel, L. Valentin, and D. Timmerman, “Comparing analytical decision support models through boolean rule extraction: A case study of ovarian tumour malignancy,” in *International Symposium on Neural Networks*, pp. 1177–1186, Springer, 2007.

[63] T. Hailesilassie, “Rule extraction algorithm for deep neural networks: A review,” *arXiv preprint arXiv:1610.05267*, 2016.

[64] R. Andrews, J. Diederich, and A. B. Tickle, “Survey and critique of techniques for extracting rules from trained artificial neural networks,” *Knowledge-based systems*, vol. 8, no. 6, pp. 373–389, 1995.

[65] T. GopiKrishna, “Evaluation of rule extraction algorithms,” *International Journal of Data Mining & Knowledge Management Process*, vol. 4, no. 3, p. 9, 2014.

[66] M. Robnik-Šikonja and I. Kononenko, “Explaining classifications for individual instances,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 5, pp. 589–600, 2008.

[67] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, “Explaining nonlinear classification decisions with deep taylor decomposition,” *Pattern Recognition*, vol. 65, pp. 211–222, 2017.

[68] S. Bach, A. Binder, K.-R. Müller, and W. Samek, “Controlling explanatory heatmap resolution and semantics via decomposition depth,” pp. 2271–2275, 2016.

[69] R. A. Berk and J. Bleich, “Statistical procedures for forecasting criminal behavior: A comparative assessment,” *Criminology & Pub. Pol’y*, vol. 12, p. 513, 2013.

[70] J. Elith, J. R. Leathwick, and T. Hastie, “A working guide to boosted regression trees,” *Journal of Animal Ecology*, vol. 77, no. 4, pp. 802–813, 2008.

[71] D. P. Green and H. L. Kern, “Modeling heterogeneous treatment effects in large-scale experiments using bayesian additive regression trees,” 2010.

[72] A. Fisher, C. Rudin, and F. Dominici, “All models are wrong but many are useful: Variable importance for black-box, proprietary, or misspecified prediction models, using model class reliance,” *arXiv preprint arXiv:1801.01489*, 2018.

[73] J. Lei, M. G’Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman,

"Distribution-free predictive inference for regression," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1094–1111, 2018.

[74] G. Casalicchio, C. Molnar, and B. Bischl, "Visualizing the feature importance for black box models," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 655–670, Springer, 2018.

[75] J. Bien, R. Tibshirani, *et al.*, "Prototype selection for interpretable classification," *The Annals of Applied Statistics*, vol. 5, no. 4, pp. 2403–2424, 2011.

[76] B. Kim, C. Rudin, and J. A. Shah, "The bayesian case model: A generative approach for case-based reasoning and prototype classification," in *Advances in Neural Information Processing Systems*, pp. 1952–1960, 2014.

[77] K. S. Gurumoorthy, A. Dhurandhar, and G. Cecchi, "Protodash: Fast interpretable prototype selection," *arXiv preprint arXiv:1707.01212*, 2017.

[78] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! criticism for interpretability," in *Advances in Neural Information Processing Systems*, pp. 2280–2288, 2016.

[79] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[80] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gpdr," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.

[81] S. Tan, R. Caruana, G. Hooker, and Y. Lou, "Detecting bias in black-box models using transparent model distillation," *arXiv preprint arXiv:1710.06169*, 2017.

[82] Z. Che, S. Purushotham, R. Khemani, and Y. Liu, “Distilling knowledge from deep networks with applications to healthcare domain,” *arXiv preprint arXiv:1512.03542*, 2015.

[83] K. Xu, D. H. Park, C. Yi, and C. Sutton, “Interpreting deep classifier by visual distillation of dark knowledge,” *arXiv preprint arXiv:1803.04042*, 2018.

[84] S. Tan, “Interpretable approaches to detect bias in black-box models,” in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 382–383, ACM, 2018.

[85] S. Tan, R. Caruana, G. Hooker, and Y. Lou, “Distill-and-compare: auditing black-box models using transparent model distillation,” pp. 303–310, 2018.

[86] S. Tan, R. Caruana, G. Hooker, and A. Gordo, “Transparent model distillation,” 01 2018.

[87] P. Cortez and M. J. Embrechts, “Using sensitivity analysis and visualization techniques to open black box data mining models,” *Information Sciences*, vol. 225, pp. 1–17, 2013.

[88] P. Cortez and M. J. Embrechts, “Opening black box data mining models using sensitivity analysis,” in *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pp. 341–348, IEEE, 2011.

[89] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, “Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation,” *Journal of Computational and Graphical Statistics*, vol. 24, no. 1, pp. 44–65, 2015.

[90] O. Bastani, C. Kim, and H. Bastani, “Interpretability via model extraction,” *arXiv preprint arXiv:1706.09773*, 2017.

[91] J. J. Thiagarajan, B. Kailkhura, P. Sattigeri, and K. N. Ramamurthy, “Treeview: Peeking into deep neural networks via feature-space partitioning,” *arXiv preprint arXiv:1611.07429*, 2016.

[92] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune, "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks," in *Advances in Neural Information Processing Systems*, pp. 3387–3395, 2016.

[93] D. Erhan, A. Courville, and Y. Bengio, "Understanding representations learned in deep architectures," *Department dInformatique et Recherche Operationnelle, University of Montreal, QC, Canada, Tech. Rep*, vol. 1355, p. 1, 2010.

[94] T. Panch, H. Mattie, and L. A. Celi, "The "inconvenient truth" about ai in healthcare," *Npj Digital Medicine*, vol. 2, no. 1, pp. 1–3, 2019.

[95] C. Ho, D. Soon, K. Caals, and J. Kapur, "Governance of automated image analysis and artificial intelligence analytics in healthcare," *Clinical radiology*, 2019.

[96] S. Qiu, Q. Liu, S. Zhou, and C. Wu, "Review of artificial intelligence adversarial attack and defense technologies," *Applied Sciences*, vol. 9, no. 5, p. 909, 2019.

[97] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pp. 506–519, ACM, 2017.

[98] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *arXiv preprint arXiv:1607.02533*, 2016.

[99] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, "The security of machine learning," *Machine Learning*, vol. 81, no. 2, pp. 121–148, 2010.

[100] D. L. Marino, C. S. Wickramasinghe, and M. Manic, "An adversarial approach for explainable ai in intrusion detection systems," in *IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society*, pp. 3237–3243, IEEE, 2018.

[101] X. Xu, X. Chen, C. Liu, A. Rohrbach, T. Darell, and D. Song, “Can you fool ai with adversarial examples on a visual turing test,” *arXiv preprint arXiv:1709.08693*, vol. 3, 2017.

[102] N. Damer, A. M. Saladié, A. Braun, and A. Kuijper, “Morgan: Recognition vulnerability and attack detectability of face morphing attacks created by generative adversarial network,” pp. 1–10, 2018.

[103] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, “Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition,” in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1528–1540, ACM, 2016.

[104] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.

[105] C. Frederickson, M. Moore, G. Dawson, and R. Polikar, “Attack strength vs. detectability dilemma in adversarial machine learning,” pp. 1–8, 2018.

[106] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” *arXiv preprint arXiv:1706.06083*, 2017.

[107] T. Gehr, M. Mirman, D. Drachsler-Cohen, P. Tsankov, S. Chaudhuri, and M. Vechev, “Ai2: Safety and robustness certification of neural networks with abstract interpretation,” *2018 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18, 2018.

[108] “Strategy for the development of quebec’s artificial intelligence ecosystem,” *A mandate from l’Économie, Science et Innovation Quebec*, 2018.

[109] J. Hare and et al, “Predicts 2019: Digital ethics, policy and governance are key to success with artificial intelligence.” URL: https://www.gartner.com/en/documents/3895092/ predicts-2019-digital-ethics-policy-and-governance-are-k, 2018.

[110] S. M. Carter, W. Rogers, K. T. Win, H. Frazer, B. Richards, and N. Houssami, "The ethical, legal and social implications of using artificial intelligence systems in breast cancer care," *The Breast*, vol. 49, pp. 25–32, 2020.

[111] B. B. Watch, "Face off the lawless growth of facial recognition in uk policing," *Obtenido de: bigbrotherwatch. org. uk/wpcontent/uploads/2018/05/Face-Off-final-digital-1. pdf. Consultado el*, vol. 22, 2018.

[112] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of ai ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019.

[113] F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. Gershman, D. O'Brien, K. Scott, S. Schieber, J. Waldo, D. Weinberger, *et al.*, "Accountability of ai under the law: The role of explanation," *arXiv preprint arXiv:1711.01134*, 2017.

[114] D. Schneeberger, K. Stöger, and A. Holzinger, "The european legal framework for medical ai," pp. 209–226, 2020.

[115] A. Alrawais, A. Alhothaily, C. Hu, and X. Cheng, "Fog computing for the internet of things: Security and privacy issues," *IEEE Internet Computing*, vol. 21, no. 2, pp. 34–42, 2017.

[116] J. Ni, A. Zhang, X. Lin, and X. S. Shen, "Security, privacy, and fairness in fog-based vehicular crowdsensing," *IEEE Communications Magazine*, vol. 55, no. 6, pp. 146–152, 2017.

[117] S. Yi, Z. Qin, and Q. Li, "Security and privacy issues of fog computing: A survey," pp. 685–695, 2015.

[118] A. Sang, R. Ismail, and C. Boyd, "A survey of trust and reputation systems for online service provision," *Decision support systems*, vol. 43, no. 2, pp. 618–644, 2007.

[119] F. Bonchi, A. Gionis, and T. Tassa, “Identity obfuscation in graphs through the information theoretic lens,” *Information Sciences*, vol. 275, pp. 232–256, 2014.

[120] R. Gennaro, C. Gentry, and B. Parno, “Non-interactive verifiable computing: Outsourcing computation to untrusted workers,” pp. 465–482, 2010.

[121] D. X. Song, D. Wagner, and A. Perrig, “Practical techniques for searches on encrypted data,” pp. 44–55, 2000.

[122] S. S. G. G. CloudWatcher, “Network security monitoring using openflow in dynamic cloud networks,” *Proc. NPSec12*, 2012.

[123] D. Balfanz, D. K. Smetters, P. Stewart, and H. C. Wong, “Talking to strangers: Authentication in ad-hoc wireless networks.,” 2002.

[124] F. Stroud, “Fog computing.” URL: https://www.webopedia.com/ TERM/F/fog-computing.html.

[125] F.Stroud,“Fogcomputing.”URL:https://www.optus.com.au/enterprise/accelerate/communications/ the-future-of-5g-enabled-edge-cloud.

[126] M. S. Hossain, G. Muhammad, and N. Guizani, “Explainable ai and mass surveillance system-based healthcare framework to combat covid-i9 like pandemics,” *IEEE Network*, vol. 34, no. 4, pp. 126–132, 2020.

[127] M. S. Hossain and G. Muhammad, “Emotion recognition using secure edge and cloud computing,” *Information Sciences*, vol. 504, pp. 589–601, 2019.

[128] P. Hu, H. Ning, T. Qiu, H. Song, Y. Wang, and X. Yao, “Security and privacy preservation scheme of face identification and resolution framework using fog computing in internet of things,” *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1143–1155, 2017.

[129] B. Sheehan, F. Murphy, M. Mullins, and C. Ryan, “Connected and autonomous vehicles: A cyber-risk classification framework,”

Transportation research part A: policy and practice, vol. 124, pp. 523–536, 2019.

[130] L. Findlater, S. Goodman, Y. Zhao, S. Azenkot, and M. Hanley, "Fairness issues in ai systems that augment sensory abilities," *ACM SIGACCESS Accessibility and Computing*, no. 125, pp. 1–1, 2020.

[131] J. Vincent, "France is using ai to check whether people are wearing masks on public transport." https:
//www.theverge.com/2020/5/7/21250357/france-masks-public-transport-mandatory-ai-surveillance-camera-so
2020.

[132] L. d. S. B. Neto, V. R. M. L. Maike, F. L. Koch, M. C. C. Baranauskas, A. de Rezende Rocha, and S. K. Goldenstein, "A wearable face recognition system built into a smartwatch and the visually impaired user.," pp. 5–12, 2015.

[133] C. L. Bennett and O. Keyes, "What is the point of fairness? disability, ai and the complexity of justice," 2019.

[134] Wikipedia, "Telephone call recording laws." URL: https://en.wikipedia.org/wiki/Telephone_call_recording\ _laws.

[135] S. Du and C. Xie, "Paradoxes of artificial intelligence in consumer markets: Ethical challenges and opportunities," *Journal of Business Research*, 2020.

[136] F. Al-Turjman, *Unmanned Aerial Vehicles in Smart Cities*. Springer, 2020.

[137] B. Dattner, T. Chamorro-Premuzic, R. Buchband, and L. Schettler, "The legal and ethical implications of using ai in hiring," *Harvard Business Review*, vol. 25, 2019.

[138] M. A. Ahad, S. Paiva, G. Tripathi, and N. Feroz, "Enabling technologies and sustainable smart cities," *Sustainable cities and society*, vol. 61, p. 102301, 2020.

[139] C. S. Lai, Y. Jia, Z. Dong, D. Wang, Y. Tao, Q. H. Lai, R. T. Wong, A. F. Zobaa, R. Wu, and L. L. Lai, "A review of technical standards for smart cities," *Clean Technologies*, vol. 2, no. 3, pp. 290–310, 2020.

[140] N. Mohammad, "A multi-tiered defense model for the security analysis of critical facilities in smart cities," *IEEE Access*, vol. 7, pp. 152585–152598, 2019.

[141] J. Huang, L. L. Presti, B. Motella, and M. Pini, "Gnss spoofing detection: Theoretical analysis and performance of the ratio test metric in open sky," *Ict Express*, vol. 2, no. 1, pp. 37–40, 2016.

[142] A. Alhilal, T. Braud, and P. Hui, "Distributed vehicular computing at the dawn of 5g: a survey," *arXiv preprint arXiv:2001.07077*, 2020.

[143] A. Musaddiq, R. Ali, R. Bajracharya, Y. A. Qadri, F. Al-Turjman, and S. W. Kim, "Trends, issues, and challenges in the domain of iot-based vehicular cloud network," pp. 49–64, 2020.

[144] Y. Agarwal, K. Jain, and O. Karabasoglu, "Smart vehicle monitoring and assistance using cloud computing in vehicular ad hoc networks," *International Journal of Transportation Science and Technology*, vol. 7, no. 1, pp. 60–73, 2018.

[145] F. Maymir, "cognitive and automatic cyber defense, nato." URL: https://www.sto.nato.int/publications/STO\%20Meeting\ %20Proceedings/STO-MP-MSG-143/MP-MSG-143-24.pdf.

[146] A. Stewart, "Three emerging innovative technologies required for cyber operations to execute commander's intent at machine speed," *Military Cyber Affairs*, vol. 4, no. 2, p. 3, 2020.

[147] D. Phillips, "The shape of cloud security in the wfh era," *ITNOW*, vol. 63, no. 1, pp. 42–43, 2021.

[148] Z. A. Collier and J. Sarkis, "The zero trust supply chain: Managing supply chain risk in the absence of trust," *International Journal of Production Research*, pp. 1–16, 2021.

[149] E. Sturzinger, K. J. Duncan, *et al.*, “Establishing and maintaining multivariate trust in a hierarchical sdn,” 2020.

[150] K. van den Bosch and A. Bronkhorst, “human-ai cooperation to benefit military decision making, nato.” URL: https://www.sto.nato. int/publications/STO\%20Meeting\%20Proceedings/ STO-MP-IST-160/MP-IST-160-S3-1.pdf.

[151] A. D’Amico, K. Whitley, D. Tesone, B. O’Brien, and E. Roth, “Achieving cyber defense situational awareness: A cognitive task analysis of information assurance analysts,” vol. 49, no. 3, pp. 229–233, 2005.

[152] J. Johnson, “Artificial intelligence & future warfare: implications for international security,” *Defense & Security Analysis*, vol. 35, no. 2, pp. 147–169, 2019.

[153] J. Johnson, “The ai-cyber nexus: implications for military escalation, deterrence and strategic stability,” *Journal of Cyber Policy*, vol. 4, no. 3, pp. 442– 460, 2019.

[154] M. Szczepanski, M. Chora´ s, M. Pawlicki, and R. Kozik, “Achieving ex-´ plainability of intrusion detection system by hybrid oracle-explainer approach,” pp. 1–8, 2020.

[155] M.-B. U. Stumbaum, *Risky business?: the EU, China and dual-use technology*. European Union institute for security studies, 2009.

[156] Y. YANG and M. AMURGIA, “How china cornered the facial recognition surveillance market.” URL: https:// www.latimes.com/business/story/2019-12-09/ china-facial-recognition-surveillance, 2020.