BRIEFING NOTES

BN-76-Emerging technology and military application-Aug2021

ETHICAL AND PRIVACY ISSUES IN AI AND IOT DEVICES

Authors: :Mehdi Taheri[1] and Kash Khorasani[2]
1 Graduate student, Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada
2 Professor, Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada

1. INTRODUCTION

1.1 AI

As opposed to the natural intelligence shown by humans, artificial intelligence (AI) is the intelligence displayed by machines. In general, intelligence is defined as cognitive problem- solving skills, which includes perceiving environment and finding analogies, calculating, and maximizing the chance of achieving goals [1]. For a machine to be intelligent, it needs to possess all the aforementioned problem-solving skills. AI was founded in 1956, and since then it has found its applications in computer science, mathematics, psychology, electrical engineering, information engineering, and many other disciplines that has made AI to become an essential component in industry and society [2].

A subset of AI in which the machines have learning capabilities and they can modify themselves when exposed to more data, is called machine learning. There are some AI algorithms that cannot be considered as machine learning, such as rules engines, expert systems, and knowledge graphs [3]. In machine learning algorithms, such as neural networks, using a historical data set, the machine attempts to minimize the error in reaching a goal by optimizing an objective function. Nowadays, generating or collecting this historical data sets and the answers that AI provides for our today's problems, have raised many concerns.

For instance, social networks have access to the personal information of millions of people and can use this information in their AI based advertisement systems to adjust the advertisement method for each person to inspire a fake need in them for a specific product. Due to the pervasive impact of AI in today's life, it is crucial to address the problems such as ethics, privacy, safety, transparency, and trust in AI. The need for ethical developments in AI has attracted many practitioners and researchers, such as IEEE initiative on Ethics of Autonomous Systems [4], the Foundation for Responsible Robotics [5], and the Partnership on AI [6].

1.2 IoT

Based on the estimations of Cisco systems, between 2008 and 2009, ratio of things or objects connected to the Internet over people connected to the Internet became more than one, and a new type of systems, called Internet of Things (IoT), was born [7]. IoT is a system of interlinked objects, animals, people, computing devices, digital systems, and mechanical devices that can communicate to each other and are able to form a network of devices [7]. This concept was described as packets of data that are transferred between different nodes, which these

nodes can be home appliances or sections of a factory, such that these nodes are integrated as an automated system [8].

There is a wide range of applications for IoT such as smart home, wearable technology, remote health monitoring, emergency notification systems, smart traffic control, vehicular communication systems, industrial IoT, Internet of Military Thing (IoMT), Internet of Battlefield Things (IoBT), and smart grids [7]. Due to the opportunities that the IoT can provide in integrating the physical world into computer-based systems, it provides us with improvement in efficiency, and economic benefits. It is estimated that by 2020, there will be 30 billion devices connected to the internet and the market value of IoT will reach $ 7.1 trillion [9, 10]. The success of the idea of connecting different devices to improve their efficiency is massively dependent on collecting, storing, and processing data. This has been done by acquisition of data from devices and storing them into a cloud network, which exposes the whole system to security and privacy problems because there is one point of vulnerability for the multiple devices.

Privacy threats in IoT, which is considered as a big data infrastructure, are of main concerns since it uses personal information of people's lives that can be used for social control or political manipulations [11]. Another major concern in adopting IoT in our life is its security. IoT devices usually have low available computation power, so that this constrain makes them unable to implement firewalls or encrypting their communications with other devices by employ strong cryptography systems.

In general, there are 4 security requirements for IOT systems: (1) data confidentiality: unauthorized access to the transmitted and stored data should be blocked; (2) data integrity: companies must detect any corruption of transmitted and stored data; (3) non-repudiation: the sender of a message should not be able to deny sending it; (4) data availability: the authorized parties should have access to the transmitted and stored data even under denial-of-service (DOS) attacks [12]. Safety is another problem that should be considered in IoT. IoT systems control actions are mainly based on event-driven smart apps that receive the data related to occurrence of an event, through a network such as Internet, from other devices and trigger the control command to an actuator [13]. Therefore, unforeseen bad app interactions, software problems, and communication failures can result in reaching dangerous and unsafe physical states [13].

1.3 AIoT

The integration of AI and IoT has led us to the emergence of Artificial Intelligence of Things (AIoT). In the AIoT system, IoT is the digital nervous system that connects different components to communicate with each other, while AI acts as the brain that makes decisions and controls the overall system [14]. Using capabilities of deep learning models and sensor telemetry data in IoT systems, AIoT can detect anomalies in real-time.

As an advantage, AIoT systems are able to proactively predict occurrence of fault in a device that may be cause a failure in the whole system. The prognosis of faults, which results in predictive and condition-based maintenance, can save millions of dollars for companies. Despite the aforementioned advantages, the AIoT system has a combination of problems that we have in AI and IoT systems, such as ethical issues, trust, transparency, privacy, safety, and security.

2. ETHICS

Recent developments in AI systems have generated an interest from researchers. One of the main questions about AI systems is, what are the moral and legal consequences of the decisions made by AI? Ethics in AI can be classified as, (1) Ethics by Design: which means ethical reasoning abilities should be part of the behavior of the systems; (2) Ethics in Design: that includes the analysis of ethical impacts of AI systems on the society; (3) Ethics for Design: the codes of conduct and standards that protects the developers and users of AI systems and ensure their integrity [15].

3. TRUST

We trust people when they explain why they are doing a specific task. Trust depends on transparency and granularity of explanations. This idea is applicable to AI systems. An AI system needs to be able to provide reasons for making a decision to users. In the digital environments, trust is called e-trust [16]. In [16], it was discussed that a bad explanation- for-trust may not create trust. For example, explain-for-trust cannot be provided by too many little details, and detailed explanation-for-confidence may not reach its goal.

As AI systems becoming more involved in making decisions for humans such as for deter- mining credit worthiness of individual and determining that whether an individual should be sentenced in a court, the importance of having trust in AI becomes more obvious [17]. There are four defined pillars of trust in [17], namely fairness, explainability, robustness, and assurance. Fairness is not achieved if we have bias in the system. Bias is essentially considered as a form of statistical discrimination that a society or population attempt to discourage [17].

Explainability is introduced in [17] as the way an individual makes derivations about an algorithm according to the level of the knowledge that the consumer has. There are three categories of explanations, first, directly interpretable by providing a companion model that represents the black box model, second, global versus local, and third, static versus interactive in which the user is able to interact with the model [17]. Assurance is related to assuring people that the AI models follow certain industry standards [17].

In [18], methods that lead us to build trust in AI systems from a regulatory point of view with a focus on the regulations in the EU have been discussed. It has been mentioned that the regulations should prevent bias and discrimination from affecting decisions made by AI. As an example, in 2017 Google was fined by the European Commission because in its shopping search comparison gave disproportionately a higher placement to its own shopping service which can be considered as a discrimination against other rival services [18].

In [18], three barriers have been identified to achieve transparency, 1) “intentional concealment on the part of corporations or other institutions”, 2) “gaps in technical literacy which, for most people, mean that having access to underlying code is insufficient”, and 3) “a lack of interpretability of the decisions made by the algorithm even to experts”. As the second set of proposed regulations in [18], entities are required to use quality labels for their products and the regulator should conduct audits and inspections of the AI companies. The third set of regulations proposed in [18] deal with the “transparency in the data chain”, which ensures that for specific decisions made by AI the data controller provides satisfactory explanation. Finally, in [18], it has been suggested to use

discrimination detection algorithms along with discrimination prevention methods subject to eliminating bias from datasets and AI algorithms.

Some metrics to measure explainability of AI systems have been provided in [19]. This paper is mainly focused on evaluation methods for the goodness of explanations, level of satisfaction of users by a given explanation, and the level of understanding of users from the AI systems. AI services providers to have satisfactory explainable AIs should pay attention to the needs of the user, the user's knowledge, and more importantly the user's goals. In [19], the explanation satisfaction is defined as the "degree to which users feel that they understand the AI system being explained to them".

The "right to explanation" was considered in the revision of the European Union's General Data Protection Regulation (GDPR) in 2018 [20]. This gives the right to users of AI systems to be informed about the existence, logic, and possible consequences of a decision made automatically by a machine [20]. Moreover, in 2016, the US Defense Advanced Re- search Projects Agency started a program regarding Explainable AI to address issues related to (1) "how to produce more explainable models"; (2) "how to design the explanation inter- face"; and (3) "how to understand the psychological requirements for effective explanations" [20]. In [20], explanation is considered as a list of abstract human-interpretable reasons or justifications that led to a certain outcome. Despite that mapping inputs and dynamics in the AI systems to human-interpretable explanations is challenging, it is feasible [20].

In [21], some methods subject to increase of trust in AI systems have been identified. It is discussed in [21] that in various stages and times in a system life cycle that create a chain of trust, four aspects should be considered, fairness, explainability, being audited, and safety.

3.1 Explainable AI

As it was mentioned earlier, to increase trust in machine the process that leads them to make a specific decision should be explainable to users. One has an

explainable AI when the machine is able to produce transparent explanations and reasons behind making a decision [22].

It is possible to achieve explainability in AI systems by using inherently explainable machine learning algorithms, such as decision trees and Bayesian classifiers [22]. As a disadvantage, employing more powerful but complicated algorithms such as neural networks results in losing transparency and explainability [22]. In an effort to provide explainability in deep learning and other complex algorithms the US Defense Advanced Research Project Agency (DARPA) has initiated a number of research projects [22]. AI explainability includes three items according to DARPA description, first, machines will explain how they reach conclusions so that future decisions can be improved, second, human users should be able to understand the decisions and trust machines. Finally, actions of AI models should be traceable and inspected by humans such that they have control over machines' decision loops [22].

In [23], three notions for various types of AI systems have been defined, namely opaque systems, interpretable systems, and comprehensible systems. Opaque systems are those AI systems in which their internal mappings from inputs to outputs are kept hidden from users. AI models which their closed-source and models are licensed by an organization and the owner wants to keep the structure of its proprietary hidden, we have opaque systems. Moreover, systems that inspection of their algorithm does not reveal their reasoning method from inputs to corresponding outputs and employ black box approaches, can be considered as opaque [23]. On the other hand, in interpretable systems users can

study and investigate

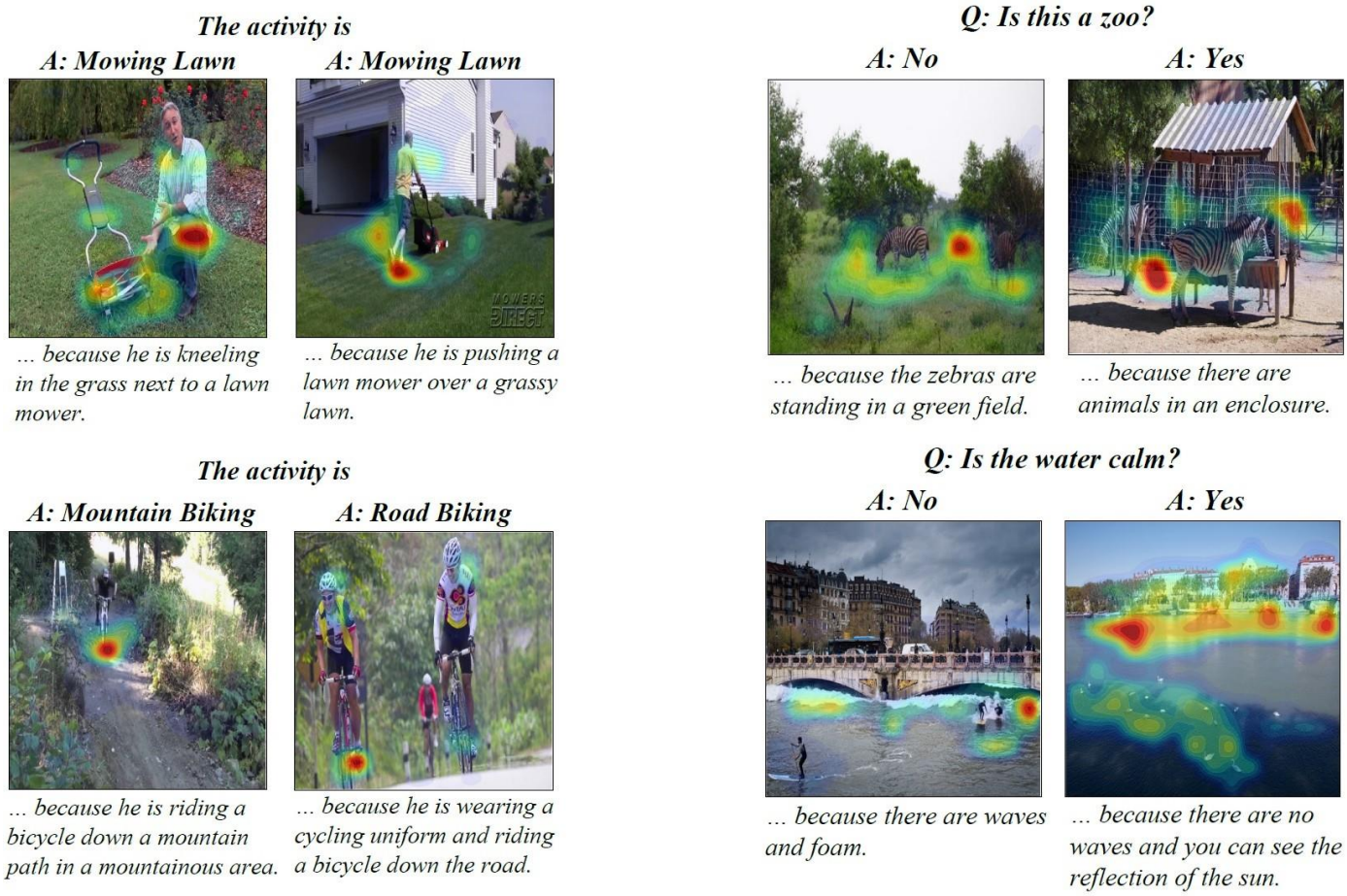

Figure 1: Examples of explainable AI [24].

the method and algorithm that are used to map a specific output from a given input. Hence, these models are transparent and users need a level of knowledge to understand details of the mapping, such as regression models and support vector machines (SVMs). But algorithms such as deep neural networks in which nonlinear transformations can be carried out cannot be considered interpretable [23]. Finally, in comprehensible systems written or visual symbols are utilized that help a user to find the relation between inputs and output. Based on the difficulty of compilation of the AI systems, one can consider different grades or levels of comprehensibility [23].

In Figure 1, four examples of explanations of an AI system for different activities and places are shown [24]. As it can be seen the upper figure on the left reasons for deciding that the activity is “Mowing Lawn”, and in the bottom figure on the left reasons for choosing “Mountain Biking” and “Road Biking” are provided. Finally, in the pictures on top and bottom on the right-side reasons for recognizing a zoo and water calm are given.

4. ACCOUNTABILITY AND TRANSPARENCY

As the autonomy in AI systems increases and they can make decisions without human control, ensuring that they have been designed ethically and responsibly, becomes more important. To achieve this, these systems should demonstrate a level of accountability and transparency. In [25], accountability is defined as the need to give reasons for and justifying the decisions made by the machine. Transparency, on the other hand, is referred to as the illustration and reproduction of the steps that have led the machine to make a specific decision and methods it uses to learn from its environment [25].

In [26], five approaches have been proposed as the main methods to address the transparency issues, namely, using simple AI models, using a combination of simple and sophisticated models, using intermediate model states, attention mechanism in which focus is on the parts of the more important input, and finally, modifying inputs such that the parts of the input that have a significant impact on the output, are chosen so that these results can be highlighted to the user. The author in [27] has proposed to establish fact sheets, similarly as in the food industry, for the AI systems. The idea of having a fact sheet for AI models was earlier proposed by IBM in [28]. IBM researchers have suggested a Supplier’s Declaration of Conformity (SDoC) that includes information about [28]:

- the type and characteristics of the service, intentions and usages of the service;
- applications that the service has been used and tested for;
- results on the performance of the service during the tests conducted by the provider and by third parties;
- giving insights on the safety of the service;
- consent of the individuals or groups that their data were used in the service and identifying the possible sources of bias;
- how the explainability is accomplished and the target user of explanation;
- the efforts were made to reduce the impact of bias, the policies that were

followed, and the methods that were used to detect bias in the service;

- the performance of the service facing unseen data or with different distributions, update in the behavior of the service in presence of newly added data, and how to monitor and test the service for drift over time;
- how the service and user's data are secured, assessment of the vulnerability of the service against attacks and its robustness, and the contingency plans in the event of adversarial attacks;
- the training datasets, the assurance on the quality of the used datasets, and the avail- ability of the datasets to the public;
- details on the methods were used to train the models, the last update of the model.

In Japan, AI R&D Principles have been introduced in the Conference of Advisory Experts of Japan's Ministry of Internal Affairs to prevent the risks in using AI, while increasing its economic and societal benefits [29, 30]. The Montreal Declaration on the Responsible Development of AI in Canada has identified seven key characteristics in developing AI, namely, "well-being, autonomy, justice, privacy, knowledge, democracy and controllability" [29, 31]. International Standards Organization (ISO) has started to set and work on essential standards that can be used to tackle safety and trustworthiness problems in AI [29].

The government of Canada has released its Directive on the Use of Machine Learning for Decision-Making that applies to any Automated Decision System from April 1, 2020 [32]. In [32], as the requirements for transparency, it has been mentioned that the service providers need to provide notice before making decisions, explain why and how their decisions were made, provide the license for components of the software, release the source code that is owned by the government of Canada and specify the access restrictions on it [32]. The European General Data Protection Regulation (GDPR), which is responsible to set rules for the purpose of data protection and protecting privacy in the EU region, has outlined the "right to explanation" that gives individuals the right to ask for an explanation when their personal data have been used by a system [29].

5. BIASING

Bias in AI mainly can occur in the stage of collecting data and the stage of preparing data [33]. The collected data can be unrepresentative of the reality, for example, a dataset that does not contain images of minorities or people of color, or it might reflect prejudices, for instance, historical employment data from the past that does not involve women in critical job positions [33]. A well-known example of biasing is the Twitter bot named Tay. Tay was designed by Microsoft to communicate with people aged 18 to 24 [34]. However, after almost 12 hours Tay became a racist bot and who twitted "should all die and burn in hell". This bot was designed to learn from the information it was programmed to receive on the internet. However, due to the differences between testing a machine in an isolated network and exposing it to a highly complex and diverse network Tay started to learn from biased information [34].

In the stage of preparing data, selecting the attributes that we want our machine to consider, which directs the behavior that the machine will have, can result in bias and dis- crimination [33]. The challenges that arise as the result of bias in AI are losing trust between humans and machines, and making decisions that express racial, gender, and ideological dis- crimination [35].

In [36], different sources of bias have been listed that can affect the fairness in AI systems such as historical bias, representation bias, measurement bias, evaluation bias, Simpsons Paradox, aggregation bias, population bias, sampling bias, and behavioral bias. It has been pointed out in [36] that fairness is achieved when there is no prejudice and discrimination. Consequently, different types of discrimination have been introduced, namely direct discrimination, indirect discrimination, systemic discrimination, statistical discrimination, explainable discrimination, and unexplainable discrimination. Moreover, various definitions of fairness each from different a point of view have been provided. For instance, fairness through awareness, equalized odds, equal opportunity, demographic parity,

5.1 How to reduce bias in AI?

The IBM Research AI group has introduced a probabilistic method for pre-processing of data subject to reducing discrimination in [37]. In this paper, a

convex optimization that can be used to learn a data transformation considering three goals "controlling discrimination, limiting distortion in individual data samples, and preserving utility" has been proposed.

One of the main problems in recognizing and reducing bias in AI systems is that there is not a general agreement on a bias metric and a fairness definition [38]. The authors in [38] have proposed a bias and fairness toolkit named "Aequitas" that can be used by both policy makers and users to evaluate machine learning models for various types of bias and fairness metrics. Policy makers, based on the application of a given AI system can choose most relevant bias metric in Aequitas to audit that system for the possible existing bias before accepting it.

In [39], fair behavior of an AI system is considered as not displaying bias or acting in that way towards any part of the population that is affected by the system's decisions. In this work, a two-step rating approach has been proposed that can be utilized to generate a scaled bias rating. The rating process should be carried out by a third party that is independent of the entity that provides the service. In this method, it is suggested in the first stage to feed unbiased input to the AI system and analyze its output. If the output is biased, the system is rated "Biased", otherwise the system should be subject to biased input in the second stage of rating. If the response of the AI system is biased, the system is labeled "Data-sensitive Biases System", and if on the contrary the output is unbiased, the system should be rated "Unbiased Compensated System" [39].

An IBM research group has introduced a toolkit that can be used for detecting and mitigating algorithmic bias in [40]. This toolkit which includes a various set of fairness metrics is named "AI Fairness 360". This toolkit contains three categories of different algorithms subject to bias mitigation, namely pre-processing algorithms in which the main effort is to transform the data subject to removing the discrimination, in-processing techniques that modify the learning algorithms to remove the bias and discrimination during the training stage. And finally, post-processing algorithms which are performed by getting access over a holdout set that was not involve the training process [36].

6. AI AND NATIONAL SECURITY

There are many significant capabilities that AI has for national security, such as cyber defense, and satellite imagery analysis, and it has been predicted that the future progress in this field will have a major impact on the strategy, organization, and priorities of the countries in the context of national security [41]. Eventually, AI will impact military superiority of countries, their information superiority, and finally, economic superiority which all of them affect national security [41]. In [41], four "lessons learned" from the past transformative military technologies, such as weaponized aircraft, have been introduced. Lesson 1, the warfare applications of AI are irresistible to be used and its military use cannot be prevented.

Hence, the goal should be to pursue safe and effective technology. Lesson 2, the commercial activities related to AI should be cultivated and restrained by the government, also policymakers need to support and protect the interests of both sides. Lesson 3, formal organizations should be created that are tasked with promoting safety by formalizing goals of technology safety. Lesson 4, the national interest of the country is influenced by the changes in the technology. In [42], the significance of threats that can occur due to international rivals in AI field has been pointed out. For instance, in 2017 the government of china revealed their plan to capture the leading position in AI by 2030. Other countries such as Russia have intentions towards development in AI as Vladimir Putin announced "whoever becomes the leader in this field will rule the world" [43].

The Chief Executive Officer of SpaceX, Elon Musk, has submitted a letter to the United Nations (UN) that warns them about the potential hazards of autonomous weapons that are controlled by AI that can "permit armed conflict to be fought at a scale greater than ever, and at timescales faster than humans comprehend" [44]. In [45], three main objectives regarding homeland security in the US have been introduced as preventing future terrorist attacks, reducing the vulnerability of the nation, and reducing the damage and recovery from attacks. The authors in [45] have provided some critical mission areas that AI can help to achieve the aforementioned objectives as follows:

- AI can contribute to recognize the patterns and activities that the attackers have to initiate warning systems and prevent attacks.
- Using image and speech recognition technology and by sharing information in the borders of countries the counter terrorism capabilities can be improved.
- AI can be used to discover cooperative relationship and patterns between criminal groups and terrorists.

- In the critical infrastructures such as water supplies, roads, and power networks, AI can be employed to detect their abnormal behaviors.
- AI can help to analyze the response plans and control the consequences of a terrorist attack.

7. FACIAL RECOGNITION

Facial recognition is a type of bio-metrics. Advances in AI, computation capabilities in machines and their memories have led to the emergence of facial recognition tools [46]. The objective of facial recognition is to have a machine that identifies a face using a camera. First, key measurements and patterns of the face, such as the distance between eyes, are evaluated by the machine. This information is stored in a database that can be updated over time. In the next step, the machines are able to capture pictures from faces and look for the possible match between its database and the captured pictures [46]. Facial recognition has applications in catching criminals, finding missing people, validation of purchases, and advertising [47]. Despite all the advantages of facial recognition technologies, they have raised concerns about making biased decisions that violate the prohibition of discrimination, ethics, privacy, and encroach on democratic freedoms [48].

7.1 Regulations to Address Bias in Facial Recognition

In the case of bias, one of the major problems in facial recognition is the high level of error in recognizing people of color and minorities. Brad Smith, the President of Microsoft, in [48] has proposed to legislate four categories of laws to address the biasing problem in facial recognition. The first category of laws should require the companies to provide their customers with transparency using some understandable documents to demonstrate the capabilities and limitations of their technology. The second category should deal with independent tests by third-parties on the facial recognition services of the companies to check the accuracy and bias in their products. The third category of laws should require the entities that provide facial recognition services to review their facial recognition outcomes by qualified people before making the final decisions. In the last category of new laws, the companies with facial recognition technology should be required to comply with and consider the laws that are in accordance with prohibiting discrimination against their costumers, in their services.

7.2 Regulations to Address Privacy Problems in Facial Recognition

Due to the widespread use of surveillance cameras around the world, the places where people visit and the pattern of their behaviors can be traced and stored easily. This information gives the governments and companies the ability to predict people's actions so that the privacy of people is violated. To avoid this violation, two types of laws are suggested in [48]. First, the entities that use facial recognition service should provide signs that clearly indicate their presence. Second, it should be mentioned in the law that entering a building or using a service that indicates the use of facial recognition in their system shows the consent of the costumers to use facial recognition.

7.3 Regulations to Protect Democratic Freedoms

Based on democratic freedoms, it is necessary for people to be able to move freely and talk to others without any governmental surveillance. Nowadays, governments are using facial recognition technology for the purpose of improving public safety. However, this ability could give governments the power to follow everyone in most public places. To address this concern, Brad Smith in [48] has proposed to have a new law which permits the governments to follow and track specified individuals using facial recognition only in the cases of having an order from a court, or in the case of emergency or immediate risk of death or injury.

8. ADVERSARIAL ATTACKS IN AI

Nowadays, sophisticated AI systems and more precisely machine learning algorithms have reached human-level performance in tasks such as image analysis, speech recognition, and natural language processing (NLP). On the other hand, despite their high level of accuracy, machine learning methodologies are vulnerable to adversarial attacks. In such attacks, inputs to the machine have been manipulated by the adversary such that their desired response is produced. Based on the adversary's knowledge of the system, in [49], attacks have been categorized as white box attacks, in which the adversary has access to the model of system, and black box attacks, where the adversary does not have direct access to the model. There are some types of attacks that fall between these two types in a sense that the adversary has a limited knowledge and access over the model of system. Moreover, in [49], based on motivations and intentions of the adversary, the adversarial attacks have been divided into four groups as follows:

- Confidentiality Attacks: In these attacks, the data that was previously used

in the training phase of AI is exposed to the malicious attacker. For example, the medical information of a high-profile politician can be obtained by a rival for possible blackmail purposes.

- Integrity Attacks: The adversary tampers the training data set of the AI such that it behaves incorrectly in response to some inputs and miss-categorizes them. These attacks can be employed to avoid spam classification and maintain the attack undetected by bypassing anomaly detection systems.
- Availability Attacks: In availability attacks, the adversary disguises its attack signals as the legitimate input to the system such that a human is not able to comprehend its differences with a healthy signal, i.e., this signal seems healthy to a human, but this compromised signal results in an incorrect output of the system. For instance, it is possible to add some noises and perturbations to the road signs that cause miss-classification by self-driving cars which can result in car crashes and dangerous situations to occur.
- Replication Attacks: This type of attacks allow the adversary to obtain a model of the system. This attack can be employed to steal intellectual property of a product.

In [50], some types of mistakes and flaws in developing machine learning models that result in a "Bad AI", such as flaws in design stage and mistakes in training phase, different types of "Malicious use of AI", methods of performing "Adversarial attacks against AI", and approaches that provide "Mitigation against adversarial attacks" have been studied. The IBM Research Ireland has released a software library named "Adversarial Robustness Toolbox" that can be used to create adversarial examples as well as defense methods for Deep Neural Networks (DNNs) [51, 52].

In [52], defending against the adversary has been divided into three stages, first, measuring model robustness by evaluating the loss of accuracy in the systems in presence of manipulated inputs by the adversary, second, model hardening by preprocessing the input subject to adding adversarial examples to the training data set, and third, runtime detection in which abnormal behaviors in the internal layers of the AI due to adversarial attacks are exploited. Automatic speech recognition (ASR) technology and NLP are being used in many devices such as cell phones, and home assistant devices to listen to the human voice and act as they are informed.

In [53], a type of adversarial examples on DNN-based ASR has been introduced which is based on "psychoacoustic hiding". In this attack on the ASR, acoustic malicious signals that are not audible by human hearing perception and contain commands to the device are successfully embedded into an arbitrary audio signal such that the device performs the adversarial tasks. The authors in [54] have studied vulnerability of Deep Q-Networks (DQNs) against perturbations and adversarial examples and introduced a novel type of attacks that provides policy manipulation in the learning phase of DQNs for the adversaries

REFERENCES

[1] Intelligence definition. [Online]. Available: http://www.brainmetrix.com/ intelligence-definition/,

[2] Artificial intelligence. [Online]. Available: https://en.wikipedia.org/wiki/Artificial intelligence

[3] Artificial intelligence (AI) vs. machine learning vs. deep learning. [Online]. Available: https://pathmind.com/wiki/ai-vs-machine-learning-vs-deep-learning

[4] [Online]. Available: https://ethicsinaction.ieee.org/

[5] [Online]. Available: http://responsiblerobotics.org/

[6] [Online]. Available: https://www.partnershiponai.org/

[7] Internet of things. [Online]. Available: https://en.wikipedia.org/wiki/Internet of things#cite note-9

[8] R. Raji, "Smart networks for control,," *IEEE Spectrum*, vol. 31, no. 6, pp. 49 – 55, 1994.

[9] A. Nordrum *et al.*, "Popular internet of things forecast of 50 billion devices by 2020 is outdated," *IEEE spectrum*, vol. 18, no. 3, 2016.

[10] C.-L. Hsu and J. C.-C. Lin, "An empirical examination of consumer adoption of internet of things services: Network externalities and concern for information privacy perspectives," *Computers in Human Behavior*, vol. 62, pp. 516–527, 2016.

[11] P. N. Howard, *Pax Technica: How the Internet of things may set us free or lock us up*. Yale University Press, 2015.

[12] S. Supriya and S. Padaki, "Data security and privacy challenges in adopting solutions for iot," in *2016 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*. IEEE, 2016, pp. 410–415.

[13] D. T. Nguyen, C. Song, Z. Qian, S. V. Krishnamurthy, E. J. Colbert, and P. McDaniel, "Iotsan: fortifying the safety of iot systems," in *Proceedings of the 14th International Conference on emerging Networking EXperiments and Technologies*. ACM, 2018, pp. 191–203.

[14] [Online]. Available: https://www.forbes.com/sites/janakirammsv/2019/08/12/ why-aiot-is-emerging-as-the-future-of-industry-40/#21ae9ce3619b

[15] V. Dignum, "Ethics in artificial intelligence: introduction to the special issue," 2018.

[16] W. Pieters, "Explanation and trust: what to tell the user in security and ai?" *Ethics and information technology*, vol. 13, no. 1, pp. 53–64, 2011.

[17] (2019) Pillars of trust in ai: Measure and quantify your ma- chine learning practices. [Online]. Available: https://medium.com/bots-and-ai/ pillars-of-trust-in-ai-measure-quantify-your-machine-learning-practices-1a5c8e72f2c6

[18] E. Thelisson, K. Padh, and E. Celis, "Regulatory mechanisms and algorithms towards trust in ai/ml," 07 2017.

[19] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable ai: Challenges and prospects," *arXiv preprint arXiv:1812.04608*, 2018.

[20] F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. Gershman, D. O'Brien, S. Schieber, J. Waldo, D. Weinberger, and A. Wood, "Accountability of ai under the law: The role of explanation," *SSRN Electronic Journal*, 11 2017.

[21] E. Toreini, M. Aitken, A. Moorsel, K. Elliott, and K. Coopamootoo, "The relationship between trust in ai and trustworthy machine learning technologies," 112019.

[22] (2019) Understanding explainable ai. [Online]. Available: https://www.forbes.com/ sites/cognitiveworld/2019/07/23/understanding-explainable-ai/#4c31a5887c9e

[23] 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.

[24] R. Goebel, A. Chander, K. Holzinger, F. Lecue, Z. Akata, S. Stumpf, P. Kieseberg, and A. Holzinger, "Explainable ai: The new 42?" in *Machine Learning and Knowledge Extraction*, A. Holzinger, P. Kieseberg, A. M. Tjoa, and E. Weippl, Eds. Cham: Springer International Publishing, 2018, pp. 295–303.

[25] V. Dignum. (2018) The ART of AI Accountability, Responsibil- ity, Transparency. [Online]. Available: https://medium.com/@virginiadignum/ the-art-of-ai-accountability-responsibility-transparency-48666ec92ea5

[26] M. Noga. (2018) Bringing transparency into AI. [Online]. Available: https://www. digitalistmag.com/future-of-work/2018/11/27/bringing-transparency-into-ai-06194523

[27] J. Rodriguez. (2018) Towards AI transparency: Four pillars required to build trust in artificial intelligence systems. [Online]. Available: https://towardsdatascience.com/ towards-ai-transparency-four-pillars-required-to-build-trust-in-artificial-intelligence-systems-d1c45a1

[28] M. Hind, S. Mehta, A. Mojsilovic, R. Nair, K. N. Ramamurthy, A. Olteanu, and K.R. Varshney, "Increasing trust in AI services through supplier's declarations of conformity," *CoRR*, vol. abs/1808.07261, 2018. [Online]. Available: http://arxiv.org/abs/1808.07261

[29] J. Millar, B. Barron, K. Hori, R. Finlay, K. Kotsuki, and I. Kerr, "Accountability in AI," in *G7 Multistakeholder Conference on Artificial Intelligence*. G7, 2018.

[30] Draft AI utilization principles. [Online]. Available: https://www.soumu.go.jp/ main content/000581310.pdf

[31] The montreal declaration for responsible AI. [Online]. Available: https://www. montrealdeclaration-responsibleai.com/

[32] (2018) Directive on the use of machine learning for decision - making. [Online]. Available: https://docs.google.com/document/ d/1LdciG-UYeokx3U7ZzRng3u4T3IHrBXXk9JddjjueQok/edit#heading=h. umd3sgrbb3d9

[33] K. Hao. (2019) This is how ai bias really happensand why its so hard to fix. [Online]. Available: https://www.technologyreview.com/s/612876/ this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/

[34] M. Garcia, "Racist in the machine: The disturbing implications of algorithmic bias," *World Policy Journal*, vol. 33, pp. 111–117, 01 2016.

[35] Many ai systems are trained using biased data. [Online]. Available: https:

//www.research.ibm.com/5-in-5/ai-and-bias/

[36] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *arXiv preprint arXiv:1908.09635*, 2019.

[37] F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *Advances in Neural Information Processing Systems*, 2017, pp. 3992–4001.

[38] P. Saleiro, B. Kuester, A. Stevens, A. Anisfeld, L. Hinkson, J. London, and R. Ghani, "Aequitas: A bias and fairness audit toolkit," 11 2018.

[39] B. Srivastava and F. Rossi, "Towards composable bias rating of ai services," 12 2018, pp. 284–289.

[40] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic *et al.*, "Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *arXiv preprint arXiv:1810.01943*, 2018.

[41] G. Allen and T. Chan, *Artificial intelligence and national security*. Belfer Center for Science and International Affairs Cambridge, MA, 2017.

[42] D. S. Hoadley and N. J. Lucas, *Artificial intelligence and national security*. Congres- sional Research Service, 2018.

[43] (2017) 'whoever leads in ai will rule the world: Putin to russian children on knowledge day. [Online]. Available: https://www.rt.com/news/401731-ai-rule-world-putin/

[44] (2017) An open letter to the united nations convention on certain conventional weapons. [Online]. Available: https://futureoflife.org/autonomous-weapons-open-letter-2017/ ?cn-reloaded=1

[45] H. Chen and F.-Y. Wang, "Guest editors' introduction: Artificial intelligence for homeland security," *IEEE intelligent systems*, vol. 20, no. 5, pp. 12–16, 2005.

[46] J. Bechtel. (2019) Two major concerns about the ethics of facial recogni- tion in public safety. [Online]. Available: https://www.designworldonline.com/ two-major-concerns-about-the-ethics-of-facial-recognition-in-public-safety/

[47] Y. Kuflinski. (2019) How ethical is facial recognition tech- nology? [Online]. Available: https://towardsdatascience.com/ how-ethical-is-facial-recognition-technology-8104db2cb81b

[48] B. Smith. (2018) Facial recognition: Its time for action. [Online]. Available: https://blogs.microsoft.com/on-the-issues/2018/12/06/facial-recognition-its-time-for-action/

[49] (2019) Adversarial attacks against ai. [Online]. Available: https://blog.f-secure.com/adversarial-attacks-against-ai/

[50] K. M. M. R. Andrew Patel, Tally Hatzakis and A. Kirichenko, "Security issues, dangers, and implications of smart information systems," 2019.

[51] (2018) The adversarial robustness toolbox: Securing ai against adversar- ial threats. [Online]. Available: https://www.ibm.com/blogs/research/2018/04/ ai-

adversarial-robustness-toolbox/

[52] M.-I. Nicolae, M. Sinn, M. N. Tran, A. Rawat, M. Wistuba, V. Zantedeschi, N. Baracaldo, B. Chen, H. Ludwig, I. M. Molloy *et al.*, "Adversarial robustness toolbox v0. 4.0," *arXiv preprint arXiv:1807.01069*, 2018.

[53] L. Schönherr, K. Kohls, S. Zeiler, T. Holz, and D. Kolossa, "Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding," *arXiv preprint arXiv:1808.05665*, 2018.

[54] V. Behzadan and A. Munir, "Vulnerability of deep reinforcement learning to policy induction attacks," in *International Conference on Machine Learning and Data Mining in Pattern Recognition*. Springer, 2017, pp. 262–275.

[55] World Wide Web Consortium and others, "The platform for privacy preferences 1.0 (P3P1. 0) specification," *World Wide Web Consortium*, 2002.

[56] (2000) Personal information protection and electronic documents act. [Online]. Available: https://laws-lois.justice.gc.ca/eng/acts/P-8.6/FullText.html

[57] (2018) Privacy commissioner issues new guidance to help address consent challenges in the digital age. [Online]. Available: https://www.priv.gc.ca/en/opc-news/ news-and-announcements/2018/nr-c 180524/

[58] (2019) Regulation of artificial intelligence: The Americas and the Caribbean. [Online]. Available: https://www.loc.gov/law/help/artificial-intelligence/americas.php

[59] Gerhard Steinke, "Data privacy approaches from US and EU perspectives," *Telematics and Informatics*, Volume 19, Issue 2, pp. 193-200, 2002.

[60] Goddard M., "The EU General Data Protection Regulation (GDPR): European Regulation that has a Global Impact," *International Journal of Market Research*, 59(6), pp. 703–705, 2017.

[61] MARGARET JACKSON, "Data Protection Regulation in Australia after 1988," *International Journal of Law and Information Technology*, Volume 5, Issue 2, SUMMER 1997, Pages 158–191.