BRIEFING NOTES

#BN-16-The role of AI-Oct2020

TRANSPARENCY, ACCOUNTABILITY, AND BIAS IN AI

Authors: Mehdi Taheri1, Reza Bahrevar[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

SUMMARY

- Artificial Intelligence (AI) has a crucial impact on today's life due to its wide range of applications in different fields, such as computer science, electrical engineering, and information engineering.
- AI systems have raised a number of concerns among both researchers and practitioners as far as issues on ethics, transparency, trust, safety, privacy, and security are concerned.
- Governments and organizations, such as Japan, Canada, IBM, and Microsoft have initiated certain regulations and initiatives to address issues such as transparency, accountability, and bias in AI.
- Lack of explainability is problematic in light of accountability and transparency of AI systems with respect to their resulting decisions for different sets of inputs.
- There are two viewpoints concerning an ethical AI. The first is moral and ethical and includes measures such as selecting unbiased data. The second is transparency of AI system's methodology.
- We argue that investing in standard development through transparency of the methodology will allow one to resolve the legal cases, as well as cyber-security issues that are related to AI systems.

CONTEXT

- As opposed to natural intelligence evidenced by humans, artificial intelligence (AI) represents intelligence displayed by machines.
- One of the main challenges in AI systems is to identify the moral and legal consequences of the decisions that are made by them. Consequently, an AI system should comply with the ethical values by its design, in its design, and for its design.
- We trust people when they explain why they are doing a specific task. This idea is applicable to AI systems. An AI system needs to be able to provide reasons for making specific decisions to the users.
- Artificial intelligence is becoming a non-separable part of emerging and disruptive technologies, and its true potential may yet to be broadly realized. Lack of explainability is problematic in light of accountability of AI systems as far as resulting decisions subject to different inputs. In particular, decisions made by these systems in applications such as autonomous driving systems, recommender systems, face recognition, speech recognition, handwriting recognition, computer vision, and automated reasoning can be questionable with regards to upholding constraints such as ethics, privacy, security, and fairness.
- There are two viewpoints concerning transparent and reliable AI systems, also referred to as ethical AI. In terms of ethical/moral grounds, the issues can be viewed in terms of what the general public and we humans believe as morally the right thing, and what steps should we

take to be proactive with regards to undesirable events such as discrimination, loss of private information, defamation, and harm to the general public/clients dealing with such systems.

- To deal with the above issue, we have to consider tools or standards such as risk assessment, mandating offline development, and data regulation to come up with policies that protect privacy and serves the interest of people.
- In terms of technical aspects, it is important to re-adjust our policies based on studies that are concerned with security and transparency of AI systems. In light of efforts made by agencies such as the European Union Agency for Cyber Security (ENISA) for developing standards for cybersecurity certifications, one needs to try to pinpoint possible elements that are necessary for development of such standards.
- Our recommendation specifically is directed at concerns that include cybersecurity as well as legal issues that are related to AI systems.
- We need to categorize the nature of threats that are posed to AI systems. We address issues regarding interpretability, robustness, and monitoring of AI systems.
- We need to define a set of measures/standards through utilization of tools such as passive security certification methods, explainable AI methods, adversarial AI, and defining passive or active observers for AI systems, to introduce criterion for achieving a safe, reliable, and transparent AI.
- Transparency requirement and using methodologies such as explainable AI, security certification, and adversarial AI implies that one should not be focusing on setting standards based on trust but rather scalable criteria. There has to be an assessable mechanism that can define trust.
- We argue that the above measures must be introduced to the willing developers as part of certification processes that allow one to understand standards through the guidelines provided to them. These certifications will be provided after carefully testing the respective AI systems using the best tools available. This will determine what level of safety they bring and what level of threat they pose to the public. Therefore first, it is clients who decide whether they are willing to use a tool having 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 potential 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 policy makers will be accountable and must work through to provide modifications for constraints that will lead to an improved, robust, and reliable AI system.

CONSIDERATIONS

- AI systems should demonstrate a level of accountability and transparency. Hence, these machines should be able to provide reasons for and justify a decision they have made and illustrate and demonstrate steps that have led them to make that decision.

- Transparency issues can be addressed by establishing fact sheets similar to the food industry for AI systems. The fact sheets should include information on the type of system, insights on its safety features, and how the system has been tested.
- We need to set and work on essential standards that can be used to tackle safety and trustworthiness problems in AI. The service providers need to provide notices before making changes and also explain why and how their decisions were made.
- Recommending standards/ measures for technical certification of AI systems.
- Possible adversarial threats towards AI.
- Discussing the interpretability of AI systems.
- Recommending methodologies/tools that can be used for settling legal cases as well as cyber-security issues related to AI systems.
- Considering the issue of accountability based on the defined standards.

NEXT STEPS *(If applicable)*

- The collected data should be representative of and correspond to all the realities. For example, a dataset must contain images of minorities or people of color otherwise it will cause bias and discrimination by the machine decision making process and lose trust between humans and machines.
- When dealing with AI systems what type of critical tools or methodologies are neglected that require allocating and dedicating more attention.
- Without proper implementation of standards such as having an AI observer, one will not be able to distinguish between the security breach or improper training of an AI system. The developer has to choose a transparent methodology and must ensure that its model is not vulnerable to adversarial threats 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 system that is ethical by design, can be assessed, and can help the government to regulate these systems.