BRIEFING NOTES

BN-60-The role of AI-May2021

ACCOUNTABILITY AND THE ROLE OF AI EXPLANATION

Authors: Mohamadreza Nematollahi[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

- Being accountable is the key for being trusted by users and society, while there should be a procedure for agencies to be able to ensure the borders of accountabilities for their products and services.
- Algorithm Impact Assessment (AIA) is a procedure that should be followed in designing any autonomous or semi-autonomous Artificial Intelligence (AI)-based software to ensure that the product or service has a net benefit for the society, while minimizing the negative impacts.
- AIA supports both the designers to ensure their responsibilities and the governments to hold them accountable subsequently. However, in case of AI systems due to their complex structure, its implementation requires that the system be designed such that its behavior is explainable.

CONTEXT

- Accountability deals with a clear acknowledgment and assumption of responsibility and answerability for actions, decisions, products, policies, and services that are rendered. It is a fundamental and crucial attribute for designers to be trusted by the society.
- Agencies should fully ensure and be aware of their responsibility borders before putting their products or services to the market and government should hold them accountable subsequently.
- Although AI systems are internally complex even for experts to understand how they will behave, this should not be an excuse for agencies using these products and designers that are designing these products to flinch from their responsibilities.

CONSIDERATIONS

- AI systems are intrinsically complex, and even for experts it is at times hard to explain and describe their internal structure that resulted in a given such behavior. This further makes the situation more challenging for agencies to ensure the accountability borders while also guaranteeing the performance of AI systems.
- The standards for agencies should properly take into account the trade secrets and should not be such that it imposes high risks of revealing trade secrets, and it should properly be addressed between designers and agencies.
- There should be proper guidelines for engaging communities and stakeholders that may be negatively affected by the consequences of AI systems in running their systems.
- AI systems can also be part of automated decision-making processes, for example, risk assessment tools in courts or any other automated decision-making processes that are utilized by various agencies. These systems represent as products under the control of agencies providing services to the public. In such circumstances, proposing mechanisms to hold those agencies accountable for social norms and fundamental democratic freedom of people and monitoring those responsibilities has a vital place in AI public

policy discussions and researches [1-2]. Nothing should prevent agencies from fulfilling their responsibilities to protect fundamental democratic values, and the complexity of the underlying processes is not a good reason to run away from those responsibilities.

- Requirements of accountability go beyond the criminal liabilities and are in the scope of the agencies' social and ethical responsibilities, even when the law may be silent regarding those issues. For example, right now in most countries, there is no law to force agencies to disclose their decision-making algorithms, and there is no well-defined mechanisms for monitoring those agencies or holding them accountable regarding fundamental democratic freedom of the people and social norms [3]. These problems should be addressed in this early development stage of AI systems, which represent the future of automatic decision-making systems.
- It is worth mentioning that there is a direct relationship between an agency's accountability and its trustworthiness by society, where an agency's social trustworthiness is a consequence of respecting their responsibilities regarding society. Without a well-defined framework for monitoring and holding the public agencies accountable and having a clear agreement of their responsibilities in the society, people and agencies will lose touch, make it impossible to detect unfairness or any other problems in the decision-making processes and making them unable to object, which consequently makes them hesitant regarding those agencies.
- Algorithm Impact Assessment (AIA) [4] defines procedures to ensure agencies' accountability scope, while considering their unwillingness to disclose their decision-making algorithms publicly. At the same time, generating explanation [5] is another method to keep the service providers and peoples which may be affected by the system in touch.
- The article [5] has tackled the role of the explanation in accountability. An explanation is a human interpretable description of the decision-making, by which a set of inputs result in a particular set of outputs. This is supposed to justify the outcomes instead of describing the internal processes that has led to those decisions. An explanation provides the answer to the following questions:
  - What are the primary and determining variables in this process? Answering this question enables the observer to determine if the right variables have been used.
  - If some of the variables can be used both in the right or wrong ways, how they have been used in this current process?
  - Why did two similar cases reach different outcomes? While the first two questions addressed the inputs' effect, this last one is mainly about the process itself.
- Explainability, on the other hand, is a property of the automated decision-making system, describes how simply an explanation can be generated. Explainability demand for simplicity, which means for a decision-making system, in which a minimum number of assumptions and input variables that are the most relevant and concrete ones have been used, generating a valid explanation would be more straightforward. It will limit the

achievable performance. Hence, Explainability is not cost-free for designers and, consequently, the users.

- Generating explanation not only is not cost-free but also may reveal trade secrets. Therefore, one should determine when an explanation would be preferable by society or is requested by the law. The following represent as situations when the benefits of explanation dominate the costs:
  - Whenever the decision has an impact on another person or any other legal personality.
  - If there is value in having an explanation, for example, if one can react based on the law, or one can request for redress.
  - The situation when one believes that an error has occurred and having evidence for the inadequacy or unreliability of the input data and in cases where one is suspicious about the output since, for example, the decision-maker gives a different output for the same cases or the same output for remarkably different cases.
  - In cases where one is suspicious about the integrity of a chain of decisions, for example, if the decisions show a remarkable benefit in favor of others.

NEXT STEPS

- Each agency aiming in using AI systems in fully autonomous or as semi-autonomous decision support systems should go through the AIA procedures leading to information that inform the public regarding the system, its purposes, policies in utilizing these systems, and the accountability borders.
- Explainability, which is a system property that enables evaluation of its decision making processes should be encouraged, since this will enable post verification of the system and facilitate implementation of the AIA procedures.
- Governments, stakeholders, and users should not over-trust the AI systems, and should actively monitor the agencies update regarding their AI utilities. For individuals to more actively be involved in, governments and public policy decision makers should educate people on these new emerging technologies.
- Nature of the decision, the susceptibility of the decision-maker to outside influence, moral and social norms, the perceived costs and benefits of an explanation, and a degree of the historical accident are also others factors which are addressed in [5], based on which the society or law request for an explanation. The agencies should be able to provide at least these explanations requested by the law and are responsible for giving a reasonable explanation when required.
- One trend is to equip the automatic decision-making systems with explanation generators, which are parallel to the central system and provide a reasonable explanation for each outcome or at least to provide some clues regarding how a specific output is related to the input or how a variable will impact the output.

- Building such system is again not cost-free and the main challenge is to build a system which is capable of generating human-interpretable explanations. In fact, explanations should be such that at least, they not create further challenges in understanding them. In cases where the financial burden of building explanation generators is not acceptable for designers, especially for small companies, alternatives are providing empirical evidences for the system behavior, or to provide theoretical guarantees for the behavior of the system, but they cannot be used in all the cases, as sometimes empirical evidences are not valid explanation, or like the big AI systems providing theoretical guarantee is not feasible.

REFERENCES

[1] https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/.

[2]https://techcrunch.com/2018/12/07/ai-desperately-needs-regulation-and-public-accountability-experts-say/.

[3] https://medium.com/@AINowInstitute/algorithmic-impact-assessments-toward-accountable-automation-in-public-agencies-bd9856e6fdde.

[4] Reisman, Dillon, et al. "Algorithmic impact assessments: A practical framework for public agency accountability." *AI Now Institute* (2018): 1-22.

[5] Doshi-Velez, Finale, et al. "Accountability of AI under the law: The role of explanation." *arXiv preprint arXiv:1711.01134* (2017).