Involving indirect stakeholders to analyse transparency in AI systems

Clàudia Figueras

Department of Computer and Systems Sciences, Stockholm University, Sweden
claudia@dsv.su.se

1 Problem statement

Accountability, responsibility and transparency have been proposed to ensure the responsible development and the incorporation of social and ethical value considerations in AI systems [3]. In this paper, we will focus on transparency. Transparency is about making clear and explicit choices and decisions concerning data sources, algorithmic design and stakeholders [3]. Useful and precise information is a prerequisite for transparency to be accessible for the users, or for the people who will be affected by the system (i.e. stakeholders). While in computer science it is common to speak of transparency in absolute terms, or as a property of the system at hand, we here approach it as a relational, dialectic and dynamic term that in practice is deeply entrenched with social and power issues.

If there is no actual defined body that takes care of what happens when harmful information is revealed, then transparency turns out to be ineffective. In this case, instead, we should focus on the power imbalance between those who are affected and those who have the power to affect [1]. As such, stakeholders should be involved in decisions about models that use personal data, affect humans or may have a morally significant impact [3] in order to balance power between decision-makers and stakeholders.

Weighing the interests of different stakeholders is the raison d'être of participatory design. In participatory design, designers, researchers and users of a technology share power and control in determining its technological future, so that the multiple stakeholders involved in the use of a technological system can have a voice in the resulting design, and the technology can better reflect their requirements, values, and concerns [4]. Concerning participatory design in AI, Lee et al. [5] developed WeBuildAI, a collective participatory framework that enables people to build an algorithmic policy for their communities, directly involving stakeholders or end-users in specifying how they would like the algorithm to behave. The authors applied WeBuildAI in a case study to a matching algorithm operating an on-demand food donation transportation service, using the input that the service's stakeholders (i.e. donors, volunteers, recipient organisations, and non-profit employees) gave to design the actual algorithm [5]. WeBuildAI framework improved procedural fairness (i.e. perceived fairness of a decision-making process), raised participants' algorithmic awareness, and helped identify inconsistencies in human decision-making in the governing organisation.

2 Research questions

AI-driven systems affect many realms of society. There is little empirical work studying how transparency engenders public trust towards organisations and AI systems [1]. Instead of focusing on the value of transparency as something that brings benefits automatically, we may instead focus on its limitations as a lens to understand accountability. As the authors state on page 984:

> "If we recognize that transparency alone cannot create accountable systems and engaging with the reasons behind this limitation, we may be able to use the limits of transparency as conceptual tools for understanding how algorithmic assemblages might be held accountable."[1]

Here, our focus is also on the practices involved in the design, development and use of an AI system; all of them configure how transparency will be perceived.

The research project will consist of exploring how different interpretations of transparency take into account the stakeholders of an AI system, particularly indirect stakeholders. With indirect stakeholders, we mean all other parties who do not directly interact with the technology but may nevertheless be affected by the use of it [4]. For example, in the case of a clinical decision support system, the direct stakeholders would be the clinicians, while the patients would be the indirect ones. The limits of transparency would be used as lenses through which to explore the transparency in such an AI system. Moreover, a sociotechnical approach, along with aspects from design justice [2] and value-sensitive design [4], are going to be included in the study.

Since rule-based and data-driven AI systems can have such an impact on society, studying those systems more in-depth, especially their direct and indirect stakeholders, can help to design a collective participatory framework, such as WeBuildAI [5], to ensure a sufficient consideration of all stakeholders' views. This work will contribute to a deeper understanding of how transparency plays a role in the interaction between stakeholders and AI systems, and in the production process, as well as how to empower indirect stakeholders.

References

1. Ananny, M., Crawford, K.: Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society **20**(3), 973–989 (Mar 2018)
2. Costanza-Chock, S.: Design Justice: towards an intersectional feminist framework for design theory and practice. Proceedings of the Design Research Society (2018)
3. Dignum, V.: Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Artificial Intelligence: Foundations, Theory, and Algorithms, Springer International Publishing (2019)
4. Friedman, B., Hendry, D.: Value Sensitive Design: Shaping Technology with Moral Imagination. The MIT Press, Cambridge, Massachusetts (2019)
5. Lee, M.K., Kusbit, D., Kahng, A., Kim, J.T., Yuan, X., Chan, A., See, D., Noothigattu, R., Lee, S., Psomas, A., Procaccia, A.D.: WeBuildAI: Participatory Framework for Algorithmic Governance. Proceedings of the ACM on Human-Computer Interaction **3**(CSCW), 1–35 (Nov 2019)

3 Mentor sessions

It would be in my most genuine interest to join the mentor sessions. Being a first-year PhD student with a topic becoming more narrowly defined opens many possible doors to configure a four-year PhD project. I would appreciate receiving additional advice on how to proceed with my research idea. Such suggestions may include examples of AI systems to study, as well as suggestions of specific research methodologies. Advice from more experienced researchers would be of my most significant benefit to succeed in my PhD.