Algorithmic Accountability for Inverse Transparency

Master's Thesis

Supervisor: Prof. Dr. Alexander Pretschner
Advisor: Valentin Zieglmeier
Email: {pretschn, zieglmev}@in.tum.de
Phone: +49 (89) 289 - 17834
Starting date: 15.11.2020

Informatik 4 – Lehrstuhl für Software und Systems Engineering
Prof. Dr. Alexander Pretschner
Fakultät für Informatik
Technische Universität München

Boltzmannstraße 3
85748 Garching bei München

https://www.in.tum.de/i04/

Context

In an increasingly digitalized world, individuals depend on technical systems that process their data. Everything from human resources to voter registrations is now handled by computer systems, which means that data are utilized. Individuals lack oversight over these systems, which can lead to discrimination and hidden biases that are hard to uncover. Recent data protection legislation tries to tackle these issues, but it is inadequate.

The research project "Inverse Transparency" tries to improve upon existing data protection by giving data owners more sovereignty in how their data are used. Its core idea is to enable access to data on a more case-by-case basis, but to monitor all accesses and make those visible to data owners. On the one hand, this can help to raise awareness of data usages and better protect the employee's personal data, on the other hand it may enable usages of data useful to teams and individuals alike.

To make this transparency meaningful, the semantics of each access need to be understood. That means explaining not just what was accessed, when, and by whom, but also what each access *means*. For analytics tools that generate insights from data based on heuristics, developers hold this knowledge and can generate meaningful log messages from their code. When it comes to machine learning and advanced so-called "algorithmic" systems, even the developers can sometimes not explain every insight generated or how it came to be.

For black box machine learning algorithms, the idea of *explainable AI* [1, 2] works towards providing a baseline of explanations. This in itself does not suffice [5, 6, 9], and it does not go beyond computer science to cover the various dimensions included in this problem. The multi-disciplinary effort of *algorithmic accountability* [10] tries to tackle how to design legislation [8], define ethical guardrails [7], and systematically design and develop computer systems [3, 4] to achieve this underlying goal.

Goal

This thesis aims to close the gap between *Inverse Transparency* and *algorithmic accountability* by developing and evaluating a systematic methodology for software design to achieve understandable transparency into algorithmic insights and decisions. The thesis has an inter-disciplinary focus, integrating ethical and psychological considerations to inform system design. The underlying assumption is that algorithmic accountability can be feasibly implemented to enable transparency into automated decisions and has the potential to improve user's sovereignty in regards to the use of their data.

Mapping study: First, methodological suggestions for and approaches to *algorithmic accountability* are surveyed in form of a systematic mapping study. The focus lies on the computer science perspective, but the results can be enriched with insights from ethics and psychology. The goal of this survey is to give an overview over relevant works, identify common themes in addressing *algorithmic accountability*, and derive a methodology with actionable requirements for system and algorithm design. This represents the first contribution of the thesis.

Implementation: Next, a messenger app with integrated analytics is designed, following the devised methodology derived from the literature. The goal is to create an exemplary prototype, so there need not be advanced algorithmic analyses. Rather, the point is to showcase how a system can offer data subjects transparency into why and how algorithmic decisions were made. This system is to represent a more extreme example of an algorithmic system, with semi- or fully automated decision making based on personal data.

The algorithmic accountability is then integrated into the Inverse Transparency toolchain. That means that when data are accessed (analyzed), the system produces an automatic explanation that is made available to the data owner.

Evaluation: Finally, a multi-step evaluation is conducted. First, the system design is deliberated from an ethical perspective considering the goals of *algorithmic accountability*, and to what

Informatik 4 – Lehrstuhl für Software und Systems Engineering
Prof. Dr. Alexander Pretschner
Fakultät für Informatik
Technische Universität München

Boltzmannstraße 3
85748 Garching bei München

https://www.in.tum.de/i04/

extent and how they are achieved. This includes comparing the implementation with messenger applications that are used in practice, and deliberating the similarities and differences.

Then, potential developers and subjects of such an algorithmic system are exposed to it and their perspectives assessed. From the perspective of the developers, the question of the feasibility of such an approach, as well as the interest in more *algorithmic accountability* is a focus.

Finally, for data subjects that might be affected by algorithmic decisions from such a system, the question of how far their (potential) concerns can be addressed by such an approach is looked at. Furthermore, they are asked to judge the quality and efficacy of the provided explanations for the chosen algorithmic decisions. Furthermore, how they would make use of the provided transparency over algorithmic decisions, assessing if that would actually increase their data sovereignty.

Work Plan

1. Conduct a systematic mapping study on approaches to *algorithmic accountability*.
2. Conceptualize a methodology for computer system design to achieve it.
3. Implement a prototype messenger app following the methodology.
4. Evaluate the methodology and prototype system theoretically and with users.
5. Document the work in the thesis.

Deliverables

- Source code of the implementation.
- Thesis written in conformance with TUM guidelines.

References

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* 6, 52138–52160.

[2] Riccardo Guidotti et al. 2019. A survey of methods for explaining black box models. *ACM Computing Surveys* 51.5, 93.

[3] Joshua A. Kroll et al. 2016. Accountable algorithms. *University of Pennsylvania Law Review* 165.3, 633–705.

[4] Bruno Lepri et al. 2018. Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology* 31.4, 611–627.

[5] Q. Vera Liao et al. 2020. Questioning the AI: Informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 1–15.

[6] Zachary C. Lipton. 2018. The mythos of model interpretability. *Queue* 16.3, 31-57.

[7] Kirsten Martin. 2019. Ethical implications and accountability of algorithms. *Journal of Business Ethics* 160.4, 835–850.

[8] Lorna McGregor et al. 2019. International human rights law as a framework for algorithmic accountability. *International & Comparative Law Quarterly* 68.2, 309–343.

[9] Brent Mittelstadt et al. 2019. Explaining explanations in AI. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*. ACM, 279–288.

[10] Maranke Wieringa. 2020. What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, 1–18.