HADI ASGHARI, NADINE BIRNER, ALJOSCHA BURCHARDT, DANIELA DICKS, JUDITH FAßBENDER, NILS FELDHUS, FREYA HEWETT, VINCENT HOFMANN*, MATTHIAS C. KETTEMANN, WOLFGANG SCHULZ, ROBERT SCHWARZENBERG, JUDITH SIMON, JAKOB STOLBERG-LARSEN, THERESA ZÜGER

*Corresponding Author, vincent.hofmann@hiig.de

What to explain when we cannot easily explain?

An interdisciplinary primer on XAI and meaningful information in automated decision-making

ABSTRACT

2021-12

WHAT TO EXPLAIN WHEN WE CANNOT EASILY EXPLAIN?

An interdisciplinary primer on XAI and meaningful information in automated decision-making

Artificial Intelligence (AI) technologies have become ubiquitous in our everyday lives, enabling new business models and intensifying the datafication of our economies. Yet, the use of AI systems entails risks on an individual as well as on a societal level. Explanations (explainable AI = XAI) can be considered a promising way to mitigate the negative effects of AI. In the following, we stick to the widely adopted term XAI. However, rather than using the notion of "AI", we refer to automated decision making (ADM) systems as the debate around explaining an automated decision is also valid for non-AI ADM systems. Explanations of the process and decision of an ADM system can empower users to legally appeal a decision, challenge developers to be aware of the negative side effects of the ADM system during the entire development process and increase the overall legitimacy of the decision. However, it remains unclear what content an explanation has to include and how the explanation can be made to achieve an actual gain of knowledge for the recipient. The GDPR provides a legal framework for explaining ADM systems. "Meaningful information about the logic involved" has to be provided to the person affected by the decision. Nonetheless, neither the text of the GDPR itself nor the commentaries on the GDPR provide details on what "meaningful information about the logic involved" precisely is.

The interdisciplinary Clinic report as well as the paper that we submit by March 2022 will approach this question from a legal, design and technical perspective. The paper proposes three questions towards a good explanation: Who needs to understand what in a given scenario? What should explanations look like in order to be *meaningful* to affected users? What do you know about the system in place to give a convincing explanation?

The outcomes could potentially not only advance the debate among legal scholars, but also help developers and designers to understand the legal obligations when developing or implementing an ADM system.

Legally, the explanation has to enable the user to appeal the decision made by the ADM system. "The logic" can be understood as "the structure and sequence of the data processing". This does not necessarily have to include a complete disclosure of the entire technical functioning of the ADM system like the score formula. Since the explanation is intended to balance the power of the ADM developer with those of the user, this balance has to be at the center of the explanation. The GDPR focuses on individual rather than collective rights. This is the subject of many discussions among scholars. However, the interpretation of the GDPR as protecting mainly individual rights is just the minimum requirement for an explanation. Any explanation going further and also having the protection of collective rights in mind, will be compliant with the GDPR as long as the individual rights are also protected. Therefore, we recommend putting the individual at the center of the explanation in a first step in order to comply with the GDPR.

With regard to the question “What should explanations look like“, we argue that XAI is more than just a technical output. To our view, XAI has to be understood as a complex communication process between human actors and cannot be merely evaluated in terms of technical accuracy. Against this backdrop, evaluating the communication process should accompany evaluating the ADM system’s technical performance. Evaluating an explanation created by an ADM system cannot be achieved without involving the user receiving the explanation. Their assessment of what a meaningful explanation needs to entail is an essential prerequisite for XAI. For domain experts, the evaluation of the explanation must include information about potentials, risks, and limitations of ADM systems for domain experts: Explainability starts even before the system is in use.

When it comes to the target group of an explanation, public or community advocates should play a bigger role. The advocate group’s interest will be more in understanding the models and their limitations as a whole instead of only focussing on the result of one individual decision.

Furthermore, transparency at the input level is a core requirement for mitigating potential bias as post-hoc interpretations are widely perceived as being too problematic to tackle the root cause. The focus should therefore shift to making the underlying rationale, design and development process transparent. For example, the use of datasheets can lead to more transparency by enabling expert users to better understand the overall process and translate it to lay users. Ultimately, using such measures will help improve ADM systems.

As we will demonstrate in both the Clinic report and the paper, there is a gap between how developers and legal experts define what explanations are. Developers aim to debug statements that help them understand their models, but these are less useful for individuals who need the explanations to be able to challenge the decision. Also, from a technical perspective, the term “logic involved” as it is used in the GDPR is at best misleading. ADM systems, and data-based systems in particular, are complex and dynamic socio-technical ecosystems. Understanding “the logic” of such diverse systems therefore requires action from different actors and at numerous stages from conception to deployment. Documenting the input data is part of the “logic involved“ from a technical perspective. In addition, developers have to explain the ADM system *how* to explain. Methods to explain the explanation often involve using additional approximate models with potentially lower accuracy. Therefore, the overall XAI process should involve direct and indirect stakeholders from the very beginning.

KEYWORDS

Artificial Intelligence, Automated-decision making, Explainability, Interdisciplinary, Internet Research