A review of the role of explanations for user acceptance in black box systems

Vinson, Norman G.; Molyneaux, Heather; Lapointe, Jean-Francois

[Proceedings of the Conference], 2018-07

A REVIEW OF THE ROLE OF EXPLANATIONS FOR USER ACCEPTANCE IN BLACK BOX SYSTEMS

Norman G. Vinson, Heather Molyneaux, Jean-François Lapointe
National Research Council, Canada
Ottawa, Ontario, Canada

ABSTRACT

The opacity of black box AI systems' decision-making has led to calls to modify these systems so they can provide explanations for their decisions. In this article we discuss what these explanations should address and what their nature should be in order to meet the concerns that have been raised and to prove satisfactory to users.

KEYWORDS

Explanation, AI, user acceptance, automated decision systems, black box.

1. INTRODUCTION

Artificial Intelligence (AI) systems are increasingly moving out of the lab into the world. As they do, the decisions these systems make will affect people's lives. AI systems can potentially make decisions about medical treatments or diagnoses, hiring and promotion, loans and the interest rates borrowers pay. Concerns have been raised about how these systems make their decisions (Guidotti et al., 2018; Wachter et al., in press). For example, in 2016 Amazon restricted some visible minority neighborhoods from participating in a free same day delivery service even though nearby areas were eligible. This lead to public outcry over alleged racial bias. Soon thereafter, Amazon extended the service to the neighborhoods in question (Ingold & Soper, 2016). Such concerns have led to calls for AI systems to explain their decisions. The question we deal with in this article is what must explanations address to be acceptable to users of AI systems and to the people who are subject to those decisions.

However, before we delve into the issue of acceptable explanations, we will provide some context by discussing the black box characteristic of many AI systems. This will lead into a discussion of the concerns that have been raised about AI decision-making. We will then discuss the issue of user acceptance of explanations.

2. BLACK BOXES

Very generally, a so-called black box system is one in which only inputs into the system and its outputs are observable (Bunge, 1963). In the context of AI, this term is also applied to systems whose input-output relationships are so complex that they cannot be understood by examining the code. Indeed, many AI systems use machine learning in which the code itself does not describe the input-output relationships but instead allows the system to build a model that encodes those relationships. This model then allows outputs to be predicted from inputs (Mohammed et al., 2017). Consequently, in such systems, examining the code cannot tell us why any particular input generates its corresponding output. Moreover, the resulting model is sometimes so complex that examining it does not provide such an understanding either.

To make matters worse, many AI decision-making systems are proprietary so that the models cannot be inspected and, in some cases, the inputs are not disclosed to protect trade secrets, privacy of the users or to ensure users do not manipulate the decision-making (Wachter et al., in press). It is for these reasons that AI decision-making systems are considered black boxes.

3. AI DECISION-MAKING CONCERNS

The black-box nature of AI systems has raised concerns that could limit adoption. The primary concern is one of fairness, in that AI decisions could be based in whole or in part on an input that should not have any influence on the decision (Zemel et al., 2013). Many jurisdictions have laws and regulations to prevent such discrimination, but it can occur with AI systems due to their black box nature (Guidotti et al., 2018).

Obviously, AI decision-making systems should refrain from using inputs that are considered discriminatory, such as race and gender. However, discrimination can creep into an AI system via the training corpus for machine learning systems (Kiritchenko & Mohammad, 2018). If the training corpus is assembled with real world cases that involve discrimination, the system will incorporate that discrimination into its predictive models. For example, if loan officers discriminated against loan applicants with foreign-looking names, an AI system model built with that data would do the same. Even if the discriminatory features are not directly input into the system, discrimination can still occur on the basis of proxy features (like location of residence) that correlate highly with the discriminatory feature (like race) (Zemel et al., 2013) This happened in the Amazon free delivery case mentioned above (Ingold & Soper, 2016).

Wachter proposed that the explanation should be in the form of a counterfactual that expresses the most similar possible world that provides the opposite decision (Wachter et al., in press). For example, a system may deny someone a loan. The corresponding counterfactual statement takes the form *<the opposite decision> would have occurred if <some condition that was not actually met had been met>*. For example, *you would have been approved for a loan if your income were 10% higher*. To limit complexity, the conditional statement should describe the closest possible world to the real world.

However, simply providing a counterfactual for each case is not sufficient to allay concerns over discrimination. Following our example, the system could discriminate on the basis of gender by requiring higher income levels for women than men. A man with a certain income could be told he would have received the loan if his income were 10% higher, but a woman with the *same* income could be told that she would need an income *15% higher* to obtain that loan. Neither that man nor that woman would know the system is discriminatory. Analyses must be conducted over *several* decisions to detect discrimination.

Contestability of a decision is also a concern (Wachter et al., in press). Decisions organizations make in regard to individuals can often be appealed. An appeal requires an understanding of *which* inputs led to the decision as well as an understanding of *how* these inputs led to the decision.

It is also important that the system be competent; that it performs well. An AI loan officer could be unbiased but if it sets low interest rates for high risk loans, the lender will lose money. To support user adoption, there would have to be evidence that the AI system has an acceptable level of performance.

Issues of fairness, contestability and competence require that explanations:

- Identify the inputs that lead to each decision,
- Show that identifiable groups of people are not intentionally or accidentally disadvantaged,
- Provide a rationale for why the identified inputs are related to the decisions,
    - Provide evidence of sufficient performance.

4. USER ACCEPTANCE

Where fairness, contestability, or competence is involved, we can argue that explanations are important. However, explanations may also be important for user satisfaction and acceptance in general (Peters, 2011).

Research on recommender system explanations can shed some light on the features of explanations that support user acceptance. Recommender systems are AI-systems that make recommendations about products or services (a book, an e-learning course) to a user, based on that user' personal characteristics or preferences. These characteristics or preferences are typically collected in a personal profile (Lapointe et al., 2017). The profile is then matched in various ways with items that are potentially of interest to the user. Interestingly, many recommender systems are not black boxes, and as a result can provide explanations for their recommendations. Researchers have studied the characteristics of recommendations that make them more acceptable to users.

Zanker (2012) found that explanations about spa recommendations increased users' perception of the system's usefulness, which increases user acceptance (Davis, 1989). The explanations revealed how the recommended items matched the users' profile. For example, one explanation could be that a recommended spa was "family friendly" for users with families. Note that it would be useful to a user to know that this was a reason for the recommendation even if the user did not intend to go to the spa with her family.

However, users have a greater preference for more complex explanations that reveal the trade-offs of various recommendations. For example, a spa may be family friendly but it may be farther away than the others. The trade-off here is family friendliness versus distance. A trade-off formulation adds other factors to a single factor explanation and supports choice-making (Pu & Chen, 2007). Indeed, users seem to prefer complex explanations over single factor explanations (Muhammad et al., 2016), especially when the issue is important to them (Pu & Chen, 2007).

Perhaps not surprisingly, people's desire for complexity is limited. Kizilcec (2016) found that people preferred more complex explanations when the results did not meet their expectations, but simpler ones when the results were expected. The added complexity involved describing the decision-making procedure. This explained why the users' expectations were violated, thus increasing user perception of fairness. However, there was a point at which additional information became confusing. For black box systems, it would be impossible (by definition) to provide a sufficiently simple and understandable explanation of the decision-making process.

Much of the research on explanations is couched in the construct of trust. Trust in electronic or computer systems is a complex construct that involves a variety of components. Some of these components, like fairness, can be addressed by explanations (Kizilcec, 2016). However, many of these components are not directly related to the characteristics of explanations provided by the system. For example, the system's reputation influences the users' trust in the system (Grabner-Krauter & Kaluscha, 2003), but reputation is not a characteristic of an explanation.

5. CONCLUSION

Many AI systems behave as black boxes, raising concerns of fairness and competence. As a result, some authors are calling for AI systems to be required to provide explanations for their decisions (Wachter, et al. in press). From a technical perspective, this poses a problem. Black box systems are not black boxes simply because their explanation feature is turned off. They are black boxes because the code does not specify which inputs lead to specific decisions and the decision-making models are typically too complex to provide sufficiently simple explanations.

Research on recommender systems that are not black boxes, and therefore can provide explanations, have revealed features of explanations that influence user acceptance. Interestingly, these features also address some of the concerns about AI decision-making raised in the literature:

- The explanation should refer to the personal characteristics of the person about whom the decision is made.
- The explanation should provide the inputs that influenced the decision (including tradeoffs).
- The explanation should describe the procedure by which inputs led to the decision, though this may not be possible for black boxes.

Since the preference for complexity seems to depend on context (Pu & Chen 2007, Kizilcec 2016), it may be preferable for people to have the opportunity to adjust how much complexity is provided by the explanation.

In addition, organizations that serve the public interest, regulators, and organizations that use AI systems themselves should want to ensure that the systems can demonstrate:

- That identifiable groups of people are not intentionally or accidentally disadvantaged.
- An acceptable level of performance.

These demonstrations require an analysis over the whole body of a system's decisions, as opposed to explanations for individual decisions. As a group, these explanations features' are far beyond what many proprietary AI decision-making systems currently provide (Wachter et al., in press). Moreover, while the concerns over AI decision-making currently focus on undesired biases, users prefer even more transparency. What exactly these preferences are, what types of explanations would be satisfactory, and how they may differ across contexts and user types remains a topic for investigation. Explanations of AI decisions is certain to be a fruitful topic for future AI and HCI research.

REFERENCES

Bunge, M., 1963. A General Black Box Theory. *Philosophy of Science*, Vol. 30, No. 4, pp. 346-358.

Davis, F. D., 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, Vol. 13, No., pp. 319–340.

Guidoitti, R. et al., 2018. A Survey of methods for explaining black box models. *arXiv preprint arXiv*:1802.01933

Grabner-Krauter, S., Kaluscha, E., 2003. Empirical Research in On-line Trust: A Review and Critical Assessment. *International Journal of Human-Computer Studies*, Vol. 58, pp. 793-812.

Ingold, D., & Soper, S., 2016. Amazon Doesn't Consider the Race of its Customers. Should it? *Bloomberg*. https://www.bloomberg.com/graphics/2016-amazon-same-day/

Kiritchenko, S., & Mohammad, S.M., 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In Proceedings of the *7th Joint Conference on Lexical and Computational Semantics (*SEM)*. New Orleans, LA, USA.

Kizilcec, R.F., 2016. How Much Information? Effects of Transparency on Trust in an Algorithmic Interface. Proceedings of the *CHI 2016 Conference ojn Human Factors in Computing Systems*, San Jose, CA, USA, pp. 2390-2395.

Lapointe, J.-F. et al. 2017. A Review of Personal Profile Features in Personalized Learning Systems. In: Andre T. (ed.) *Advances in Human Factors in Training, Education, and Learning Sciences (AHFE 2017); Advances in Intelligent Systems and Computing*, Vol. 596, pp. 46-55, Springer.

Mohammed, M. et al., 2017. *Machine Learning: Algorithms and Applications*. CRC Press, New York, USA.

Muhammad, K. et al., 2016. A Live-User Study of Opinionated Explanations for Recommender Systems. *In proceedings of IUI'16, 21st International Conference on Intelligent User Interfaces*, Sonoma, CA, USA pp. 256-260.

Peters, W., 2011. Explanation and trust: what to tell the user in security and AI? *Ethics and Information Technology*, Vol. 13, Issue 1, pp. 53-64.

Pu, P., & Chen, L., 2007. Trust-Inspiring Explanation Interfaces for Recommender Systems. *Knowledge-Based Systems*, Vol. 20, Issue 6, pp. 542-556.

Wachter, S. et al., in press. Counterfactual explanations without opening the black box: Automated decisions and the GDPR, *Harvard Journal of Law & Technology*.

Zanker, M. 2012. The Influence of Knowledgeable Explanations on Users' Perception of a Recommender System. Proceedings of the sixth *ACM Conference on Recommender Systems RecSys'12*, Dublin Ireland, pp. 269-272.

Zemel, R. et al., 2013. Learning Fair Representations. In Proceedings of the *30th International Conference on Machine Learning (ICML 2013)*, Atlanta, Georgia, USA. *Journal of Machine Learning Research, Workshops and Conference Proceedings (JMLR: W&CP)*, Vol. 28, No. 3, pp. 325-333.