# Do explanations for data-based predictions actually increase users' trust in AI?

5 October 2020, by Ingrid Fadelli

Credit: Photos Hobby, Unsplash

In recent years, many artificial intelligence (AI) and robotics researchers have been trying to develop systems that can provide explanations for their actions or predictions. The idea behind their work is that as AI systems become more widespread, explaining why they act in particular ways or why they made certain predictions could increase transparency and consequently users' trust in them.

Researchers at Bretagne Atlantique Research Center in Rennes and the French National Center for Scientific Research in Toulouse have recently carried out a study that explores and questions this assumption, with the hope of better understanding how AI explainability may actually impact users' trust in AI. Their paper, published in *Nature Machine Intelligence*, argues that an AI system's explanations might not actually be as truthful or transparent as some users assume them to be.

"This paper originates from our desire to explore an intuitive gap," Erwan Le Merrer and Gilles Trédan, two of the researchers who carried out the study, told TechXplore. "As interacting humans, we are used to not always trusting provided explanations, yet as computer scientists, we continuously hear that explainability is paramount for the acceptance of AI by the general public. While we recognize the benefits of AI explainability in some contexts (e.g., an AI designer operating on a 'white box'), we wanted to make a point on its limits from the user (i.e., 'black box') perspective."

Many researchers have recently argued that machine learning algorithms and other AI tools should be able to explain the rationale behind their decisions, in a similar way to humans. Le Merrer and Trédan, on the other hand, believe that while AI explanations might have value in local contexts, for instance providing useful feedback to developers who are trying to debug a system, they might be deceptive in remote contexts, where an AI system is trained and managed by a specific service provider, thus its decisions are delivered to users via a third party.

"A user's understanding of the decisions she faces is a core societal problem for the adoption of AI-based algorithmic decisions," Le Merrer and Trédan explained. "We exposed that logical explanations from a provider can always be prone to attacks (i.e., lies), that are difficult or impossible to detect for an isolated user. We show that the space of features and possible attacks is very large, so that even if users collude to spot the problem, these lies remain difficult to detect."

To better explain the reasoning behind their ideas, Le Merrer and Trédan drew an analogy with bouncers outside nightclubs, who might lie when explaining to individual customers why they are denied entry at the door. Similarly, the researchers suggest that remote service providers could lie to users about the reasoning behind an AI's predictions or actions, for instance, to leverage discriminative features. In their paper, they refer to this parallelism as "the bouncer problem."

"Our work questions the widespread belief that explanations will increase users' trust in AI systems," Le Merrer and Trédan said. "We rather conclude the opposite: From a user perspective and without pre-existing trust, explanations can easily be lies and therefore can explain anything anyhow. We believe that users' trust should be sought using other approaches (e.g., in-premises white box algorithm auditing, cryptographic approaches, etc.)."

In their paper, Le Merrer and Trédan provide practical examples of how the explainability of AI actions in remote contexts could be affected by "the bouncer problem." In the future, their work could inspire further studies exploring the benefits and limitations of developing machine learning algorithms or robots that can explain the reasoning behind their actions, while also potentially prompting the development of alternative solutions for increasing people's trust in AI.

"We plan to continue studying AI systems from the users' (i.e., 'black box') perspective, particularly exploring this question: What can regular users discover/learn/understand/infer about the AI systems that shape a growing part of their life?" Le Merrer and Trédan said. "For instance, we are currently studying the phenomenon of user shadow banning (i.e., blocking or partially excluding a user from being able to reach an online community) on platforms that claim that they are not using this method."

**More information:** Erwan Le Merrer et al. Remote explainability faces the bouncer problem, *Nature Machine Intelligence* (2020). DOI: 10.1038/s42256-020-0216-z



APA citation: Do explanations for data-based predictions actually increase users' trust in AI? (2020, October 5) retrieved 25 November 2020 from https://techxplore.com/news/2020-10-explanations-data-based-users-ai.html