Explain to whom? Putting the User in the Center of Explainable AI

Alexandra Kirsch
[orcid.org/0000−0002−5663−1798]

Universität Tübingen
alexandra.kirsch@uni-tuebingen.de

Abstract. The ability to explain actions and decisions is often regarded as a basic ingredient of cognitive systems. But when researchers propose methods for making AI systems understandable, users are usually not involved or even mentioned. However, the purpose is to make people willing to accept the decision of a machine or to be better able to interact with it. Therefore, I argue that the evaluation of explanations must involve some form of user testing.

1 Reasons for Explanations

I regularly present the following definition by Brachman [1] in my AI lecture:

> A truly cognitive system would be able to learn from its experience — as well as by being instructed — and perform better on day two than it did on day one. It would be able to explain what it was doing and why it was doing it. It would be reflective enough to know when it was heading down a blind alley or when it needed to ask for information that it simply couldn't get to by further reasoning. [...]

Once a student raised his hand and asked whether the ability to explain one's own actions is a necessary feature of cognition. After all, he cannot explain all his actions (and if he does, it is doubtful whether such an explanation reflects the psychological processes that really led to the action) and still would be considered as a cognitive system.

The ability of an agent to explain its own decisions is often regarded as a necessary feature without a clear understanding of why agents should have this ability. Connected to this is the question of what an explanation is. With the current hype of statistical machine learning, the general feeling is that these methods need some way to make their decisions plausible to people (like the DARPA initiative Explainable Artificial Intelligence[1]). But when researchers talk about explanations, they hardly ever consider the end user. In this paper I argue that comprehensibility and explainability must always be regarded in the context of a specific use case and that its adequacy can only be determined by interaction with users.

[1] https://www.darpa.mil/program/explainable-artificial-intelligence

2 Usability Principles

The idea to build comprehensible technical devices is much older than the current trend in AI. Many usability guidelines from the fields of design and human-computer interaction include implicit or explicit explanation to users. Norman [2] proposes four basic design principles, all of them contain elements of explainability:

conceptual model: users should understand the underlying mechanisms of the system;
visibility: all functionality should be visible, a kind of explanation of what the system is capable of;
mapping: the visible elements should intuitively map to functionality;
feedback: the user should always be informed about the system's state.

Thus, any device should be designed in a way that users are aware of the system's capabilities and current state. This can be interpreted as a kind of explanation, one can also say that such a system does not need any additional explanation. For AI researchers this means that 1) comprehensible systems can often be built by adhering to well-known usability principles, 2) the design and usability of the overall system is often more important than AI features to result in an overall positive user experience.

3 Legibility

Embodied and virtual agents differ from typical human-computer interaction in that they can perform more pronounced actions and often have additional sensing capabilities. What is an explanation in this context? Of course, such agents could generate explanations of their actions, possibly in natural language. But in many cases a constant stream of explanations may annoy users and its effectiveness is questionable. When we look at human-human interaction, we hardly ever need to explain our actions to others. So why should a machine do so?

A more natural way of interaction could be to implicitly communicate goals and the necessity of actions. Lichtenthäler and Kirsch [3] define the term *legibility* as follows:

> Robot behavior is legible if: (Factor 1) a human observer or interactor is able to understand its intentions, and (Factor 2) the behavior met the expectations of the human observer or interactor.

This definition centers on the human observer, implying that legibility can only be determined by experiments involving users. Such tests can again be inspired by standard usability testing, but the quality criteria and setup must often be adapted to the task and embodied situation of the agent. In addition, if physical robots are involved, there are additional questions of how to ensure the safety of participants.

As an example, consider the task of legible robot navigation [4]. Robot navigation has traditionally been treated as a purely technical problem, but with humans sharing the space of the robot or directly interacting with it, legibility becomes an issue. But then, the quality of a navigation act cannot just be characterized by its success and possibly the time needed to reach the goal point, one must also determine whether human observers can indeed infer the goal point and that the human expectations are met [5].

Performing user tests is time-consuming. Therefore, it would be nice to have a general set of measures that determine legibility, but that can be measured without direct user involvement. An experiment of Papenmeier et al. [6] identifies two factors that influence the legibility of robot navigation: 1) the movement direction: a robot looking towards the direction it is moving to (or towards the goal point, both were identical in the experiment) is perceived as more autonomous than one with a different orientation while moving; 2) the velocity profile: deceleration of the robot causes surprise in human observers (as measured by a viewing time paradigm), while acceleration and constant velocities do not.

4 Explainability and Interaction

Legibility as used above is a criterion for passive interaction: even though a person has no direct contact with the agent, the behavior should be legible. But it is also a prerequisite for direct interaction, for example a person trying to initiate an interaction would have to be able to predict the robot's movements. But direct interaction may require more than just a basic, implicit understanding of the other's intentions, it may at times require a more explicit explanation.

Decision methods can be designed in a way that resembles human decision making. The Heuristic Problem Solver [7] and the FORR architecture [8] follow closely the way that humans make decisions: they explicitly generate alternatives and evaluate them. In this way, an explanation is generated along with the solution: the alternatives that were considered as well as the reasons for choosing one can directly be communicated to users [9].

In addition, such an explicit decision-making paragidm enables an interactive process of humans and machines jointly determining a solution. It is practically impossible to model all the knowledge a person has with respect to a task. But if people can interact with the decision process of the machine, they can use this additional knowledge without having to formalize all of it. Thus, a person could propose or delete alternatives, or change the evaluations of alternatives.

Here again, the question is whether the explanation is adequate. Ideally, the machine should provide enough information for a person to be satisfied with or even be willing to take responsibility for the decision. Or in the case of interaction, a possible measurement is whether the joint decision-making is more efficient and effective than a purely computational or purely human one.

5 Conclusion

The ability to explain one's decisions and actions is often unquestioningly required from cognitive systems. But the frequency, form and implementation of explanations depend on the specific application and context. In many cases, standard feedback mechanisms known from human-computer interaction are all that is needed. Other applications may demand that users are able to identify the source of errors. Even others may need an explicit dialogue between human and machine that allows the machine to request help if necessary. Explanations can take different forms, from standard interaction methods of flashing lights, sounds and graphical displays, to language or the display of legible actions. In the last section I have mentioned systems that generate an explanation along with the solution. An alternative is the post-hoc generation of explanations for black box algorithms, which is currently popular with the rise of statistical machine learning.

In all respects, the users are the leveling rule. An explanation is not a mathematical construct, an explanation is good if people find it helpful in the specific context. Therefore, the AI community should expand its evaluation metrics beyond optimization criteria to user-centered measures and evaluation procedures.

.

References

1. Brachman, R.: Systems that know what they're doing. IEEE Intelligent Systems (November/December 2002) 67 – 71
2. Norman, D.A.: The Design of Everyday Things. Basic Books, Inc., New York, NY, USA (2002)
3. Lichtenthäler, C., Kirsch, A.: Legibility of Robot Behavior : A Literature Review. preprint at https://hal.archives-ouvertes.fr/hal-01306977 (April 2016)
4. Kruse, T., Pandey, A.K., Alami, R., Kirsch, A.: Human-aware robot navigation: A survey. Robotics and Autonomous Systems **61**(12) (2013) 1726–1743
5. Lichtenthäler, C., Lorenz, T., Karg, M., Kirsch, A.: Increasing perceived value between human and robots — measuring legibility in human aware navigation. In: IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO). (2012) 89–94
6. Papenmeier, F., Uhrig, M., Kirsch, A.: Human understanding of robot motion: The role of velocity and orientation. preprint at https://osf.io/tphnu/ (2017)
7. Kirsch, A.: Heuristic decision-making for human-aware navigation in domestic environments. In: 2nd Global Conference on Artificial Intelligence (GCAI). (2016)
8. Epstein, S., Aroor, A., Evanusa, M., Sklar, E., Simon, S.: Navigation with learned spatial affordances. (2015)
9. Korpan, R., Epstein, S.L., Aroor, A., Dekel, G.: WHY: Natural explanations from a robot navigator. In: AAAI 2017 Fall Symposium on Natural Communication for Human-Robot Collaboration. (2017)