How to Achieve Explainability and Transparency in Human AI Interaction

Joana Hois(✉), Dimitra Theofanou-Fuelbier, and Alischa Janine Junk

Group Research, Future Technologies, Daimler AG, Stuttgart, Germany
{joana.hois,dimitra.theofanou-fuelbier,
alischa_janine.junk}@daimler.com

Abstract. It is typically not transparent to end-users, how AI systems derive information or make decisions. This becomes crucial, the more pervasive AI systems enter human daily lives, the more they influence automated decision-making, and the more people rely on them. We present work in progress on explainability to support transparency in human AI interaction. In this paper, we discuss methods and research findings on categorizations of user types, system scope and limits, situational context, and changes over time. Based on these different dimensions and their range and combinations, we aim at individual facets of transparency that address a specific situation best. The approach is human-centered to provide adequate explanations with regard to their depth of detail and level of information, and we outline the different dimensions of this complex task.

Keywords: Transparency · Explainability · Human AI interaction

1 Introduction

The number of artificial intelligence applications that are available on the business and consumer market have increased over the last years (Das et al. 2015). In some areas, more tasks have even been taken over by intelligent algorithms. Also, the future impact of AI is expected to become further pervasive and encompassing. One such example is a lifelong personal assistant (Gil and Selman 2019) that supports and tutors humans. These systems will highly affect social lives and influence human decisions. Trusting and relying on such systems to make correct (or 'good') suggestions or decisions is inevitable for these AI systems to achieve their full functionality (Mohseni et al. 2018).

AI systems can provide explanations together with their decisions and suggestions or interact with users when questions about their decisions and suggestions arise: in human-computer-interaction – or rather human-AI-interaction – explainability provides transparency and contributes to trust (Miller 2019). Even though trust itself is influenced by a variety of other aspects, e.g., human, robotic, and environmental factors (Schaefer et al. 2016), we focus here on aspects regarding explainability when interacting with artificial intelligence systems and how this can yield transparency. Also, the need for explainability of AI systems' decisions and behaviors has grown in general (Gunning 2017), and explainability is seen as a toolset to understand the underlying technicalities and models (Ribeiro et al. 2016 and Štrumbelj and Kononenko 2014).

C. Stephanidis (Ed.): HCII 2019, CCIS 1033, pp. 177–183, 2019.
https://doi.org/10.1007/978-3-030-23528-4_25

For more adaptive, continuously learning AI systems that closely collaborate with human end-users and that may change their behavior over time, transparency and understanding of the AI systems' behavior is inevitably, e.g., to increase user acceptance. The exact way, how to achieve this transparency and explainability is still an open question and ongoing research shows the complexity of the entire topic (Miller 2019 and Mohseni et al. 2018). For example, users may vary the detail of transparency they wish to see, or users may react more seamlessly to the system's behavior with higher understanding.

In this paper, we discuss different levels of transparency both from the perspective of human end-users and AI systems. In the next section, we show the different dimensions of transparency both from a human and an AI perspective. We next address potential roles and relationships during the human-AI-interaction, followed by aspects of situational awareness and time. As a result, we highlight the complexity when aiming at an appropriate level of explanations with regard to transparency in a specific situation.

2 Facets of Transparency

The existing body of research concerned with transparency and explainability of AI focus on different aspects of transparency, see Sect. 3. In this paper we will use a three-facetted-model of transparency based on the work of (Endsley 1995) and (Chen et al. 2014) regarding the situation awareness model and agent transparency.

As shown in Fig. 1, we identify three key facets of transparency. One aspect being the transparency about the behavior and the underlying intentions of the system. The second facet is concerned with the decision making mechanism of the system, including an understanding about the underlying algorithm and the integrated variables. The third facet adds an understanding about potential limitations of the system which includes an estimation of the probability of errors in a given situation.

When determining the level of transparency in a given situation, characteristics of the system as well as the user have to be taken into account: the system can provide explanations actively or on-demand and the system can also interact in a specific way that may be interpreted as social cues by the user. The user has certain preferences and prior experience with systems and potential expectations. A facet of transparency can be achieved by the interaction of both system and user.

The adequacy of an explanation, however, can hardly be determined without taking into account personal characteristics of the user. Depending on the general technical knowledge, the time of usage and the situation awareness of the user, the required quality and quantity of the explanation to reach a certain level of transparency might vary. This effects possible relationships during interaction (Sect. 4) and is influenced by specific situations (Sect. 5).

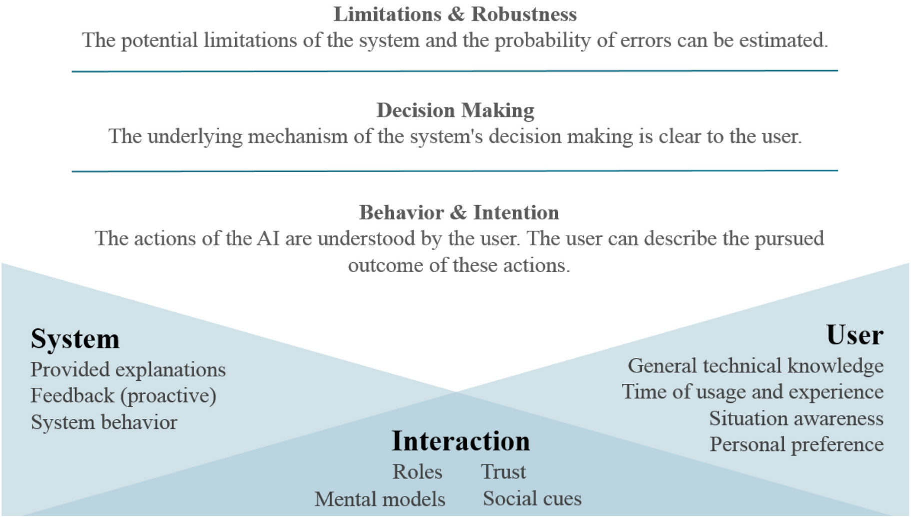
Limitations & Robustness
The potential limitations of the system and the probability of errors can be estimated.
Decision Making
The underlying mechanism of the system's decision making is clear to the user.
Behavior & Intention
The actions of the AI are understood by the user. The user can describe the pursued outcome of these actions.
System
Provided explanations
Feedback (proactive)
System behavior
User
General technical knowledge
Time of usage and experience
Situation awareness
Personal preference
Interaction
Roles
Trust
Mental models
Social cues

Fig. 1. Facets of transparency

3 Aspects of AI Explainability

AI functionalities are nowadays often enabled by machine learning models that have been trained with large data sets and that may learn when interacting with users and change their behavior over time. It has been argued that certain models intrinsically entail explanations in their decisions, e.g., decision trees, and are thus more easy to interpret, though also decision trees can become rather complex for humans to perceive and understand them (Štrumbelj and Kononenko 2014 and Došilovic et al. 2018). Complex machine learning models are difficult to interpret, and several approaches for explainability have been discussed (Ribeiro et al. 2016 and Samek et al. 2017).

It can be distinguished, whether explainability is primarily seen as a method that aims at analyzing trained machine learning model results or as a method that aims at making machine learning model results transparent for end-users. Analyzing an AI system according to all aspects is recommended (Mohseni et al. 2018). In this paper, we focus on those aspects directly related to interaction with end-users, who are not experts in technical details or the developers of the AI system.

Explainability of AI systems have several different aspects:

- The system can use different channels to communicate explanations, such as text, speech, graphical, visualizations, or auditive signals.
- Measures to evaluate explainability for non-expert users vary between measuring user mental models, task performance, user satisfaction, or trust, according to (Mohseni et al. 2018).
- The main purpose to provide explainability of a model also varies, e.g., the goal might be to support trust, causality, transferability, informativeness, or ethical reasons, according to (Lipton 2016).

- Finally, the exact content that is used to communicate explanations can be distinguished. This might depend on context and situation, user-specific preferences, or technical likelihoods. In short, a user might prefer a short but easy to understand explanation over an elaborate but difficult to comprehend explanation. The meta level can also vary, e.g., a system communicates its decision making, its technical aspects, its limitations, or options for alternative decisions, cf. (Miller 2019).

The different dimensions that have to be considered for a transparent human AI interaction are shown in Table 1. To adequately address all dimensions in a specific situation, an AI system thus requires different options to select, which information to provide for an explanations, which depth of detail, and when to provide explanations. End-users might have a higher need for detailed explanations when confronted with unexpected AI decisions than for routine decisions. However, further aspects are relevant as presented in the next sections.

Table 1. Aspects of explainability in AI systems for end-user interaction

Content	Channel	Evaluation	Purpose
Detailed	Visual (text)	Mental model	User acceptance
Brief	Visual (graphics)	Task performance	Trust
Individualized	Auditive	User satisfaction	Causality
General	Movement	Trust measures	Information
…	…	…	…

4 Relations Between Humans and AI During Interaction

(Fitts 1951) characterized the human-machine interaction by describing the relative strengths and limitations of humans and computers, sometimes referred to as what "men are better at" and what "machines are better at" lists (MABA-MABA). Since the classification includes the full range between "only human" and "only machine", a description of different levels of automation (LOA) became necessary, e.g. (Sheridan and Verplank 1978, Parasuraman et al. 2000), see Table 2. Despite the wide body of research in the field of LOA of the last 60 years, the question of how the human decision making process could be implemented in autonomous systems has not been answered yet. While systems with integrated machine learning algorithms are developed, that are able to learn and change their behavior over time, the situation becomes even more complex. E.g., while a certain limitation of a system (e.g., sensor fusion) might lead to the presentation of the full set of decision alternatives at the beginning, it might change over time to the next higher level of automation where only one alternative is suggested. A different facet of transparency (see Sect. 2) might be needed to ensure a suitable interaction after a certain time of usage.

When interacting with an intelligent systems, yet another aspect comes into play: the attribution of roles, such as the AI system being a tutor or a personal assistant. Further research will have to clarify, if different roles of the intelligent system might have implications for the recommended level of automation, action selection, and

Table 2. Levels of automation of decision and action selection (Parasuraman et al. 2000, p. 287)

HIGH
10. The computer decides everything, acts autonomously, ignoring the human.
9. informs the human only if it, the computer, decides to
8. informs the human only if asked, or
7. executes automatically, then necessarily informs the human, and
6. allows the human a restricted time to veto before automatic execution, or
5. executes that suggestion if the human approves, or
4. suggests one alternative
3. narrows the selection down to a few, or
2. The computer offers a complete set of decision/action alternatives, or
1. The computer offers no assistance: human must take all decisions and actions.
LOW

transparency. (Karapanos et al. 2009) has shown that human expectations towards a product changes over time. In terms of a personal intelligent assistant, for instance, this may also be applicable, and the way a human perceives and interacts with an intelligent system may shift over time as the user makes experiences with the system.

5 Situational Awareness and Context

As argued before, the personal characteristics of the user as well as the characteristics of the system have an impact on the recommended type of explanation and the interaction quality. Additionally, the context in which the interaction takes place is expected to have a significant influence on the interaction in general and the need for explanation and transparency in particular. The situation awareness of the user and the time of usage are key factors to influence the need for transparency and explanation in order to create trust.

According to (Endsley 1995), situation awareness encompasses the perception of the situation, the comparison of the situation, and the anticipation of a future state. In this paper, the term situation awareness will be used to refer to the characteristics of the situation as well as possible consequences of the decision making. The relationship between the situation awareness and the need for transparency and explanation however is not linear.

The situation characteristics further impacts the trust level a user places in the AI system or its explanation. Studies have shown that explanations can increase trust or the lack of explanation can decrease trust, e.g., (Holliday et al. 2016). Trust aspects are more relevant though, when dealing with severe situations. Particularly, when situational awareness is rather low, trust becomes more relevant (Wagner and Robinette 2015). On the one hand, humans may still trust and rely on systems making poor decisions (Wagner and Robinette 2015). Ideally in these situations of overtrust, a system would be able recognize its own limitations and make it transparent. On the other hand, humans also tend to disbelieve explanations given by an already untrusted systems (Miller 2019).

6 Summary and Outlook

An intelligent system that aims at making its behavior, decisions, and suggestions transparent to human users in a specific situation has to take into account various facets and dimensions, as described above. In this paper, we highlighted the various topics that lead to the complexity of such an endeavor.

Further research is needed with regard to long-term studies that show how the interaction between learning systems and users may change over time and thus vary with regard to transparency. In this respect, the impact of trust and changes in trust with the support of transparency is also an open topic.

Furthermore, transparency is not only complex and cost- or time-expensive, its wide variations with regard to a specific situation is particularly influenced by consequences of the interaction. Routine situations may not rely on transparency, while severe situations heavily depend on it. Transparency could also be offered after interaction has taken place, e.g., the situation and the underlying mechanisms how decisions were made by the system could be presented to the user after a critical situation. Such adequate ways, however, need to be studied.

Personality traits could be of interest for a situation-adequate human AI interaction: users with a need for cognition might have a higher need for explanations or technically averse users may need additional explanations. However, in severe situations, this might not be as relevant.

References

Chen, J.Y.C., Procci, K., Boyce, M., Wright, J., Garcia, A., Barnes, M.: Situation awareness–based agent transparency. Technical report, Army Research Laboratory ARL-TR-6905 (2014)

Das, S., Dey, A., Pal, A., Roy, N.: Applications of artificial intelligence in machine learning: review and prospect. Int. J. Comput. Appl. **115**(9), 31–41 (2015)

Došilović, F.K., Brčić, M., Hlupić, N.: Explainable artificial intelligence: a survey. In: Proceedings of the 41st International Convention on Information and Communication Technology, Electronics and Microelectronics MIPRO, pp. 210–215. IEEE Xplore (2018)

Endsley, M.R.: Toward a theory of situation awareness in dynamic systems. Hum. Factors J. **37** (1), 32–64 (1995)

Fitts, P.M.: Human engineering for an effective air navigation and traffic control system. Technical report, National Research Council (1951)

Gunning, D.: Explainable artificial intelligence (XAI). DARPA Program (2017). https://www.darpa.mil/program/explainable-artificial-intelligence. Accessed 18 Mar 2019

Gil, Y., Selman, B.: A 20-year community roadmap for artificial intelligence research in the US executive summary. https://cra.org/ccc/wp-content/uploads/sites/2/2019/03/AI_Roadmap_Exec_Summary-FINAL-.pdf. Accessed 18 Mar 2019

Holliday, D., Wilson, S., Stumpf, S.: User trust in intelligent systems: a journey over time. In: Proceedings of the 21st International Conference on Intelligent User Interfaces, pp. 164–168. ACM, New York (2016)

Karapanos, E., Zimmerman, J., Forlizzi, J., Martens, J.-B.: User experience over time: an initial framework. In: CHI 2009 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 729–738. ACM, New York (2019)

Lipton, Z.C.: The mythos of model interpretability. Commun. ACM **61**(10), 36–43 (2016)

Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. J. **267**, 1–38 (2019)

Mohseni, S., Zarei, N., Ragan, E.D.: A survey of evaluation methods and measures for interpretable machine learning. Computing Research Repository (CoRR) (2018). http://arxiv.org/abs/1811.11839. Accessed 18 Mar 2019

Parasuraman, R., Sheridan, T.B., Wickens, C.D.: Model for types and levels of human interaction with automation. IEEE Trans. Syst. Man Cybern. – Part A: Syst. Hum. **30**, 286–297 (2000)

Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should i trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144. ACM, New York (2016)

Samek, W., Wiegand, T., Müller, K.-R.: Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. ITU J.: ICT Discov. Impact Artif. Intell. (AI) Commun. Netw. Serv. **1**(1), 39–48 (2017)

Schaefer, K.E., Chen, J.Y.C., Szalma, J.L., Hancock, P.A.: A meta-analysis of factors influencing the development of trust in automation: implications for understanding autonomy in future systems. Hum. Factors: J. Hum. Factors Ergon. Soc. **58**(3), 377–400 (2016)

Sheridan, T.B., Verplank, W.: Human and computer control of undersea teleoperators. Man-Machine Systems Laboratory, Department of Mechanical Engineering, MIT, USA (1978)

Štrumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst. **41**(3), 647–665 (2014)

Wagner, A., Robinette, P.: Towards robots that trust: human subject validation of the situational conditions for trust. Interact. Stud. **16**(1), 89–117 (2015)