Considerations on Explainable AI and Users' Mental Models

Heleen Rutjes
h.rutjes@tue.nl
Eindhoven University of Technology
Eindhoven, The Netherlands

Martijn C. Willemsen
m.c.willemsen@tue.nl
Eindhoven University of Technology
Eindhoven, The Netherlands

Wijnand A. IJsselsteijn
w.a.ijsselsteijn@tue.nl
Eindhoven University of Technology
Eindhoven, The Netherlands

ABSTRACT

As the aim of explaining is understanding, XAI is successful when the user has a good understanding of the AI system. This paper shows, using theories from the social sciences and HCI, that appropriately capturing and accounting for the user's mental model while explaining is key to successful XAI.

CCS CONCEPTS

• **Human-centered computing** → **User models**; *HCI theory, concepts and models*; • **Computing methodologies** → **Theory of mind**; *Cognitive science.*

KEYWORDS

Explainable AI, theory of mind, mental models, user models, mathematics education

ACM Reference Format:
Heleen Rutjes, Martijn C. Willemsen, and Wijnand A. IJsselsteijn. 2019. Considerations on Explainable AI and Users' Mental Models. In *Where is the Human? Bridging the Gap Between AI and HCI, Workshop at CHI'19, May 4–9, 2019, Glasgow, Scotland Uk*. ACM, New York, NY, USA, 5 pages. https://doi.org/0

CHI 2019, May 4–9, 2019, Glasgow, Scotland Uk

ACM ISBN 0…$0
https://doi.org/0

INTRODUCTION

As Artificial Intelligent (AI) systems are increasingly prevalent in our daily lives, our interactions with these systems are of frequent occurrence, e.g., when we receive movie or music recommendations, when the decision to be invited for a job interview is based an algorithm, or when our doctor uses a medical decision support system to diagnose a disease. As a consequence, in addition to focusing on the technological challenges, AI research is increasingly taking a human-centered approach to the design of AI. More specifically, research on explainable AI (XAI) focusses on how these systems should communicate with users, when it comes to explaining and justifying the outcomes of an AI model.

In the social sciences, there is a vast and valuable body of research that is potentially relevant to understand the user's needs in XAI. However, currently research on transparency and explainability of AI has failed to consider insights from cognitive psychology [1] or social sciences in general [11]. Even the field of Human-Computer Interaction (HCI) has not been actively involved in the research and development of AI systems, until recently [8].

Learning, and thus explaining, is not a straightforward task. This holds for human teachers and their students, but even more for XAI, because of the different nature of humans and computers. Computers handle explicit knowledge better than tacit knowledge [3], and perform best in specific and well-defined tasks, that are low in uncertainty [4]. At the same time, people's behaviors and experiences are highly contextualized, ambiguous and social, and it is debated whether it is at all possible for computers to "understand" its users to their full extent [10, 15]. However, full understanding of the user, through complete and accurate user modelling, might also not be required for effective XAI. What level of user modelling is possible, and more importantly, what level of user modelling is sufficient for effective XAI, is still an open research question.

In this paper, we take a closer look into potentially relevant literature from the social sciences and HCI, and argue that, for XAI to be effective, it should at least accurately capture and consider the user's mental model of the AI system while generating explanations. A mental model is a person's internal representation of the people, objects and environments she is interacting with. The mental model, ideally, builds a correct understanding of the way the world (or, in our case, the AI system) works, allowing for comprehension of the current state of affairs, and prediction of future states. Systems of high complexity typically require explanations in order to achieve accurate mental models, which assist in effective, efficient and pleasurable use of those systems.

USERS' MENTAL MODELS OF AI SYSTEMS

Miller [11] highlights the social nature of explanations, and the importance of a system's *model of self* and a *theory of mind* of the user for successful explainability. The system's *model of self*, that may be an approximation of the true model, reasons about its own model to generate explanations. The

theory of mind keeps track of information shared with the user so far, and thus forms a simple model about the user's knowledge, understanding and beliefs. Miller [11] states: "If it has such a model, the explanatory agent can exploit this by tailoring the explanation to the human observer." (p.57)

This *theory of mind* resembles the *user's mental model* in Norman's framework on human-centered design [13, 14], which to date is still a highly influential framework in HCI. The framework describes that a designer generates a system image, alongside the system, that is an explanation of the system, e.g., an interface or a manual. The mechanisms and capabilities of the system are explained to the user through this system image. The user constructs her own mental model of the system through this system image, and by interacting with the system. Norman [14] summarizes: "Good conceptual models are the key to understandable, enjoyable products: good communication is the key to good conceptual models." (p.32). This idea is fundamental to XAI: to make full and enjoyable use of AI systems, it is of key importance that the user understands the system well.

The importance of sufficient consideration of a user's mental model, is also reflected in research on mathematics education, where similarly, explanations of abstract concepts are studied. For example, the Van Hiele model [7] describes five phases a student goes through while learning a geometrical concept, starting from a superficial and visual level, evolving into more abstract and rigorous levels of understanding. Van Hiele [7] highlights that effective teaching requires adjustment to the level of the student, and "that this [different levels of] reasoning calls for different languages" (p.252). Furthermore, Tall and Vinner [16] differentiate between a (mathematical) concept and a (student's) *concept image*. They emphasize that, whereas mathematics is built on formal logic, the human brain trying to grasp these mathematical concepts is not an entirely logical entity. Therefore, students might build erroneous concept images, not accurately reflecting the mathematical concept, which may seriously hinder learning, especially when entering more advanced levels. The authors highlight the importance of taking the student's concept image in consideration for effective teaching. Yet, they note it may be hard for teachers to identify such erroneous concept images, arguing for a teaching process where the student's concept images are regularly and thoroughly checked.

These theories highlight not only the importance, but also the challenge teachers and designers are facing, to carefully adjust explanations to the level of understanding of students (or users). It shows that the user's mental model of an abstract concept is likely to be different than the technical and mathematical properties of the concept itself. Those mental models may be wrong, e.g., incomplete or imprecise [6]. In case they conflict with the true properties of the system, it may seriously hinder the user's understanding, even when adequate explanations are given. At the same time, Greca and Moreira [6] highlight the function of mental models, not necessarily related to the accuracy: "The main role of a mental model is to allow its builder to explain and make predictions about the physical system represented by it. It has to be functional to the person who constructs it." (p.3). In that perspective, mental models might not be *wrong*, but *different*, in a way that they are construed on a different

level of understanding than the explanatory agent assumes, or originate from different definitions of success, rooted in different contexts or beliefs. One might even argue that concepts as fairness, accuracy and good explainability of AI systems are social constructs, which can only be defined in coordination with all stakeholders involved: the AI system and the user. In either case, for both wrong and different mental models, starting from the user's perspective, and closely following her mental models throughout the explanation process, is crucial for successful XAI.

Some studies do take a human-centered approach in understanding and defining fairness [2] and explainability [5] of AI systems, but this is rather the exception than the rule. Many current explanations of AI are driven by knowledge from computer science rather than social science [12], often resulting in explanations of specific aspects, e.g. classifiers or factor weights, rather than explanations that address a user's information needs or anxieties. In order to provide the user with appropriate explanations, we might need to know what triggered her to interact with the system anyway, what her goals, prior knowledge and beliefs are, whether she is under time-pressure, perhaps even anxious of loss of face when she does not immediately understand, and so on, and so forth.

Lastly, users' mental models can exist at several levels, and thus their corresponding explanations exist at different levels. For example, on a low operational level, users may have an understanding of how and why a specific picture is classified as an animal. On a higher level, users may have mental models on the capabilities and trustworthiness of the system. On this level, there can be a lot at stake. For example, if a user's mental model overestimates the intelligence of a system because the systems has a 'human-like' appearance, this may result in over-reliance of the system and lowers the user experience [9]. Or, if users' mental models underestimate the control that they have over a system, they can draw the incorrect conclusion that the system will replace rather than augment them. In this light, explanations can play a vital role in trust-calibration and technology adoption.

CONCLUSION

Concluding, as the final aim of explaining is understanding, XAI should focus on good user's understanding. To achieve this, it is of key importance to appropriately capture a user's mental model and account for this while generating explanations. There is a vast and valuable body of research in the social sciences and HCI on how users construct their mental models, which can be of great guidance to make XAI more effective.

ACKNOWLEDGMENTS

Thanks to John Lokman, for challenging and shaping our ideas.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI '18)*. 1–18. https://doi.org/10.1145/3173574.3174156

[2] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. https://doi.org/10.1145/3173574.3173951 arXiv:1801.10408

[3] Harry Collins. 2010. *Tacit and Explicit Knowledge*. The University of Chicago Press, Chicago and London. https://doi.org/10.1188/12.CJON.341-342

[4] Mary Cummings. 2014. Man versus Machine or Man + Machine ? *IEEE Computer Society* 29, 5 (2014), 62–69.

[5] Motahhare Eslami, Sneha R Krishna Kumaran, Christian Sandvig, and Karrie Karahalios. 2018. Communicating Algorithmic Process in Online Behavioral Advertising. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)* (2018). https://doi.org/10.1145/3173574.3174006

[6] Ileana María Greca and Marco Antonio Moreira. 2000. Mental models, conceptual models, and modelling. *International Journal of Science Education* 22, 1 (2000), 1–11. https://doi.org/10.1080/095006900289976

[7] Pierre M. Van Hiele. 1959. Perspective of The Childs Thought and Geometry. In *English translation of selected writings of Dina van Hiele-Geldof and Pierre M. van Hiele*. Chapter 6, 243–252.

[8] Kori Inkpen, Munmun De Choudhury, Stevie Chancellor, Michael Veale, and Eric P.S. Baumer. 2019. Where is the Human? Bridging the Gap Between AI and HCI. https://ai-hci.github.io/

[9] Bart P. Knijnenburg and Martijn C. Willemsen. 2016. Inferring Capabilities of Intelligent Agents from Their External Traits. *ACM Transactions on Interactive Intelligent Systems* 6, 4 (2016), 1–25. https://doi.org/10.1145/2963106

[10] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The Parable of Google Flu: Traps in Big Data Analysis. *Science* 343, March (2014), 1203–1206. https://doi.org/10.1126/science.1248506

[11] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007 arXiv:1706.07269

[12] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of Inmates Running the Asylum. Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. In *IJCAI-17 Workshop on Explainable AI (XAI)*. 36–42. https://doi.org/10.1016/j.foodchem.2017.11.091 arXiv:1712.00547

[13] Don A. Norman. 1983. Some observations on Mental Models. In *Mental Models*, Dedre Gentner and Albert L. Stevens (Eds.). Psychology Press, Chapter 1, 7–14.

[14] Don A. Norman. 2013. *The design of everyday things*. Basic books, New York. https://doi.org/10.1002/hfm.20127 arXiv:arXiv:1011.1669v3

[15] Heleen Rutjes, Martijn C. Willemsen, and Wijnand A. IJsselsteijn. 2019. Beyond Behavior: The Coach's Perspective on Technology in Health Coaching. In *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI '19)*.

[16] David Tall and Shlomo Vinner. 1981. Concept image and concept definition in mathematics with particular reference to limits and continuity. *Educational Studies in Mathematics* 12, 2 (1981), 151–169. https://doi.org/10.1007/BF00305619