4-17-2022

Logic and Pragmatics in AI Explanation (Chapter)

Chun-Hua Tsai

John M. Carroll

Logic and Pragmatics in AI Explanation

Chun-Hua Tsai[1(✉)] and John M. Carroll[2]

[1] University of Nebraska at Omaha, Omaha, NE, USA
chunhuatsai@unomaha.edu
[2] Pennsylvania State University, University Park, PA, USA
jmc56@psu.edu

Abstract. This paper reviews logical approaches and challenges raised for explaining AI. We discuss the issues of presenting explanations as accurate computational models that users cannot understand or use. Then, we introduce pragmatic approaches that consider explanation a sort of speech act that commits to felicity conditions, including intelligibility, trustworthiness, and usefulness to the users. We argue Explainable AI (XAI) is more than a matter of accurate and complete computational explanation, that it requires pragmatics to address the issues it seeks to address. At the end of this paper, we draw a historical analogy to usability. This term was understood logically and pragmatically, but that has evolved empirically through time to become more prosperous and more functional.

Keywords: Explainable AI · Pragmatics · Conversation · Causability

1 Introduction

Artificial intelligence (AI) technology has advanced many human-facing applications in our daily lives. As one of the most widely used AI-driven intelligent systems, recommendation systems have been an essential part of today's digital ecosystems. For example, recommendation systems have been widely adopted for suggesting relevant items or people to the users on social media [8]. Billion people have adopted or interacted with these AI systems every day. Effective recommender systems typically exploit multiple data sources and ensemble intelligent inference methods, e.g., machine learning or data science approaches. However, it is usually difficult to comprehend the internal processes of how the recommendation was made for the end-users. The *reasons* of receiving specific recommendations usually stay in a *black box*, which frequently makes the resulting recommendations less *trustworthy* to the users [1]. The users generally have little understanding of the mechanism behind these systems, so these recommendations are not yet transparent to the users. The opaque designs are known to negatively affect users' satisfaction and impair their trust in the recommendation systems [25]. Moreover, in this situation, processing this output could produce user behavior that can be confusing, frustrating, or even dangerous in life-changing scenarios [1].

A. Holzinger et al. (Eds.): xxAI 2020, LNAI 13200, pp. 387–396, 2022.
https://doi.org/10.1007/978-3-031-04083-2_19

We argue providing explainable recommendation models and interfaces may not assure the users will *understand* the underlying rationale, data, and logic [26]. The *scientific explanations*, which are based on accurate AI models, might not *comprehensible* to the users who are lack competent AI literacy. For instance, a software engineer would appreciate inspecting the approximated probability in a recommendation model. However, this information could be less meaningful or even overloaded to lay users with varied computational knowledge, beliefs, and even biases [2]. We believe the nature of an explanation is to help the users to understand and to build a working mental model of using AI applications in everyday lives [5]. We urgently need more work on empowering lay users by providing comprehensible explanations in AI applications to benefit from the daily collaboration with AI.

In this paper, we aim to review *logical* approaches to Explainable AI (XAI). We would review the logic of explanation and challenges raised for explaining AI using generic algorithms. Specifically, we are interested in presenting such explanations to users, for instance, explaining accurate system models that users cannot understand or use. Then, we would discuss pragmatic approaches that consider explanation a sort of speech act that commits to felicity conditions, including intelligibility, trustworthiness, and usefulness to the listener. We argue XAI is more than a matter of accurate and complete explanation, that it requires pragmatics of explanation to address the issues it seeks to address. We then draw a historical analogy to usability. This term was understood logically and pragmatically, but that has evolved empirically through time to become more prosperous and functional.

2 The Logic of Explanations

Explainable AI (XAI) has drawn more and more attention in the broader field of human-computer interaction (HCI) due to the extensive social impact. With the popularity of AI-powered systems, it is imperative to provide users with effective and practical transparency. For instance, the newly initiated European Union's General Data Protection Regulation (GDPR) requires the owner of any data-driven application to maintain a "right to the explanation" of algorithmic decisions [7]. Enhancing transparency in AI systems has been studied in the XAI research to improve AI systems' explainability, interpretability, or controllability [14,16]. Researchers have explored a range of user interfaces and explainable models to support exploring, understanding, explaining, and controlling recommendations [10,25,26]. In many user-centered evaluations, these explanations positively contribute to the user experience, i.e., trust, understandability, and satisfaction [25]. Self-explainable recommender systems have been proved to increase user perception of system transparency and acceptance of the system suggestions [14]. These explanations were usually post-hoc and one-shot with an obvious challenge of when, why, and how to explain the system to the users based on their information needs and beliefs.

Another stream of research has identified the effects of making the recommendation process more transparent. It could improve the user's conceptual

model by enhancing the recommendation system's controllability [22,26]. In these attempts, users were allowed to *influence* the presented recommendations by interacting with different visual interfaces. The interactive recommender systems demonstrated that users appreciate *controllability* in their interactions with the recommender systems [14]. The similar effects applied to visualization that users can understand how their actions can impact the system, which contributes to the overall *inspectability* [14] and *causability* [13] of the recommendation process. The transparent recommendation process could accelerate the information-seeking process but does not guarantee the comprehension of the target system's inner logic. These solutions empowered the user to control the system for accessing the desired recommendations. However, these controllable interfaces may not fulfill the explanation needs and help the users build a mental model to tell how the system works.

The user's mental model represents the knowledge of information systems generated and evolved through the interaction with the system [18]. The idea was founded in cognitive science and HCI discipline in the 1980s. For instance, Norman [21] argued the user could invent a mental model to simulate system behavior and make assumptions or predictions about the interaction outcome based on a target system. Follow Norman's definition, the user's mental models are constructed, incomplete, limited, unstable and sometime "superstitions" [21]. The user's mental model interacts with the *conceptual model* that the system designer used to develop the system. HCI researchers have considered the user's mental model in designing the usable system or interfaces in the past two decades. However, only a few studies have examined the user's mental model while interacting with the context of AI-powered recommender systems and algorithmic decisions [20].

We argue that these controllable and explainable user interfaces may not always ensure that users understand the underlying rationale of each contributing data or method [26]. The users could perceive the system's usefulness but still lack the *predictability* or *causability* [13] that to approximate the behavior of the target system [21]. In our observation, the users could build different mental models while interacting with an explainable system. For instance, users with more robust domain knowledge, such as trained computer sciences students, would be more judgmental in using the explainable system through their computational knowledge. However, the *naive* users would be more willing to accept and trust the recommendations [26]. We also observe controllable interfaces would lead the user to *compare* the recommendations in their decision-making process. Still, it does not mean the users could understand or predict the system's underlying logic. These findings demonstrate that personal factors and mental models (such as education, domain experience, and familiarity with technology) could significantly affect the system's user perception and cognitive process of machine-generated explanations.

3 The Pragmatics of Explanations

Miller [18] and Mittelstadt et al. [19] suggest that the AI and HCI researchers need to differentiate *scientific* and *everyday* explanations. To provide the *everyday* explanations, researchers need to consider cross-discipline knowledge (e.g., HCI, social science, cognitive science, psychology, etc.) and the user's mental model. Instead of the scientific intuition to provide prediction approximations (e.g., the global or local surrogate XAI models). For example, as HCI researchers, we already know the success explanation should be iterative, sound, complete, and not overwhelm the user. Social science researchers defined the everyday explanation through three principles [24]. 1) *human explanations are contrastive*: perceiving abnormality played an important role in seeking an explanation, i.e., the users would be more like to figure out an unexpected recommendation [18]. 2) *human explanations are selective*: the users may not seek a "complete cause" of an event; instead, the users tend to seek useful information in the given context. The selective could reduce long causal chains' effort and the cognitive load of processing countless modern AI models' parameters. 3) *human explanations are social*: the process of seeking an explanation should be interactive, such as a conversation. The explainer and explained can engage in information transfer through dialogue or other means [12].

Specifically, we propose to explore the *pragmatics of Explanations in AI*, i.e., the known mechanism of how the user requests an explanation from AI applications. The HCI community has long been interested in the interaction benefits of conversational interfaces. The design space could be situated within a rich body of studies on conversational agents or chatbot applications, e.g., AI-driven personal assistant [17]. The design of conversational agents offers several advantages over traditional WIMP (Windows, Icons, Menus, and Pointers) interfaces. The interface could provide a natural and familiar way for users to tell the system about themselves, which improves the system's usability and updates the user's mental model to the system. Moreover, the design is flexible (like a dialogue) and can accommodate diverse user requests without requiring users to follow a fixed path (e.g., the controllable interfaces [26]). The interaction could augment by a personified persona, in which the anthropomorphic features could help attract user attention and gain user trust.

In this section, we present two case studies to introduce our early investigation on pragmatics of AI explanations.

3.1 Case 1: Conversational Explanations

Online symptom checkers (OSCs) are intelligent systems using machine learning approaches (e.g., clinical decision tree) to help patients with self-diagnosis or self-triage [27]. These systems have been widely used in various health contexts, e.g., patients could use OSCs to check their early symptoms. The patient could learn their symptoms before a doctor visit, and to identify the appropriate care level and services and whether they need medical attention from healthcare providers [23]. The AI-powered symptom checkers promise various benefits,

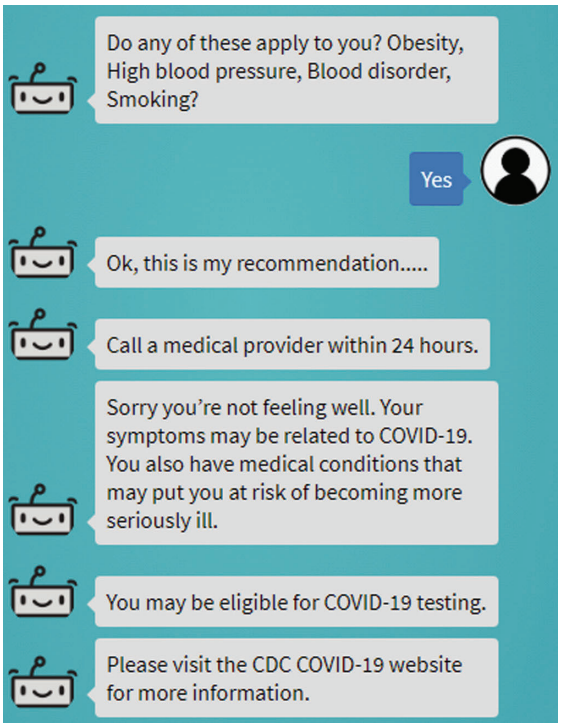

Fig. 1. Example of the conversational AI explanations [27]

such as providing quality diagnosis and reducing unnecessary visits and tests. However, unlike real healthcare professionals, most OSCs do not *explain why* the OSCs provide such diagnosis or *why* a patient falls into a disease classification. OSCs' data and clinical decision models are usually neither transparent nor comprehensible to lay users.

We argue *explanations* could be used to promote diagnostic transparency of online symptom checkers in a conversational manner. First, we conducted an interview study to explore *what explanation needs exist in the existing use of OSCs*. Second, informed by the first study's results, we used a design study to investigate *how explanations affect the user perception and user experience with OSCs*. We designed an COVID-19 OSC (shown in Fig. 1) and tested it with three styles of explanations in a lab-controlled study with 20 subjects. We found that conversational explanations can significantly improve overall user experiences of trust, transparency perception, and learning. Besides, we showed that by interweaving explanations into conversation flow, OSC could facilitate users' comprehension of the diagnostics in a dynamic and timely fashion.

The findings contributed empirical insights into user experiences with explanations in healthcare AI applications. Second, we derived conceptual insights into OSC transparency. Third, we proposed design implications for improving transparency in healthcare technologies, and especially explanation design in conversational agents.

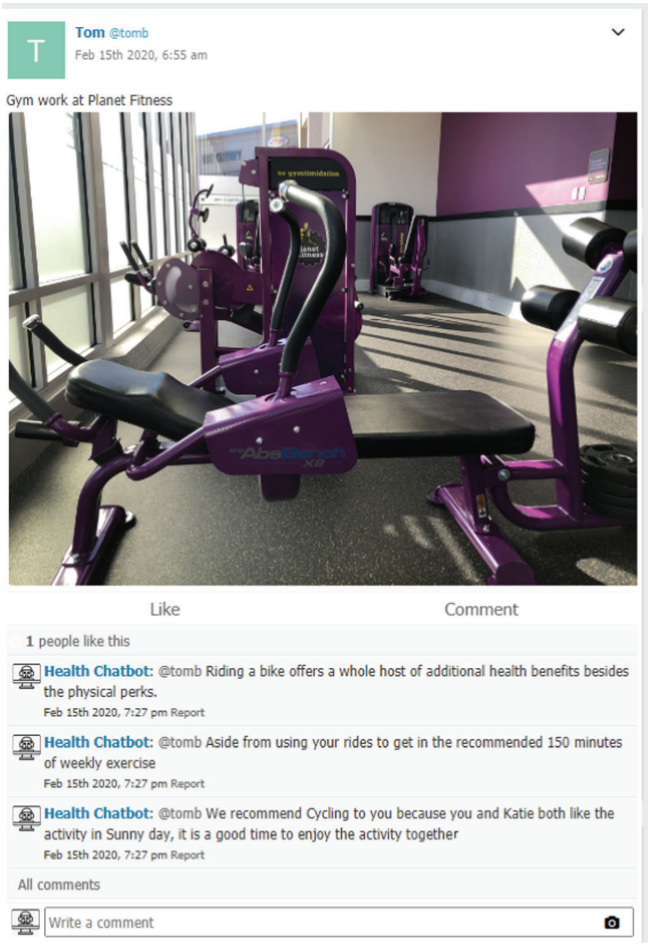

Fig. 2. Example of the Explainable AI-Mediated communication.

3.2 Case 2: Explainable AI-Mediated Communication (XAI-MC)

The integral part of modern health promotion initiatives for non-collocated members is computer-mediated communications [11]. The concept has been extensively adopted as an interpersonal communication medium in public health research such as telemedicine and mental health supports. Today, the Artificial Intelligence-Mediated Communication (AI-MC) between people could be augmented by computational agents to achieve different communication goals [9]. For instance, the interpersonal text-based communications (e.g., email) could be augmented by auto-correct, auto-completion, or auto-response. AI-MC has received more and more attention in recent socially efficacious research. For example, an AI agent could undermine the writers' message by altering the negative keywords (e.g., "sorry") to encourage the user to normalize language as the

right way of speaking. AI agent could mitigate interpersonal biases by triggering alert messages when the agents detected the users intend to post negative messages on social media [15]. The introduction of AI brings new opportunities to adopt computational agents in family health collaboration and communications. AI-MC could be used to engage family members' health conversation better, and the communication may translate into healthy behavioral changes [6]. We can introduce a designate agent to mediate the communication by *recommending and explaining the health information* to the family members. Little attention has been paid to the question of how computational agents ought to disclose to users in AI-MC and the effects on family health promotion.

We explored the effects of promoting non-collocated family members' healthy lifestyle through *Explainable AI-meditated Communication (XAI-MC).* We examined how XAI-MC would help non-collocated family members to engage in conversations about health, to learn more about each other's healthy practices, and as a result to encourage family collaboration via an online platform. We are particularly interested in exploring the effect of bringing transparent AI agents to the family communication. Specifically, we proposed to design a transparent AI agent to mediate the non-collocated family members' communication on healthy lifestyles. In our design, the users could share healthy activities information for enhancing family health awareness and engagement in a social media application. In the platform (shown in Fig. 2), a designate AI-powered health chat bot was used to mediate family members' communication on social media by explaining the health recommendations to them. We adopted the explainable health recommendations to address existing challenges related to remote family collaboration on health through XAI-MC. The findings could help to generate insights into designing transparent AI agents to support collaborating and sharing health and well-being information with online conversation.

We conducted a week-long field study with 26 participants who have at least one non-collocated family member or friend willing to join the study together. Based on a within-subject design, participants were assigned to two study phases: *1) AI-MC with non-explainable health recommendation* and *2) XAI-MC with explainable health recommendation.* We adopted a mixed-method to evaluate our design by collecting quantitative and qualitative feedback. We found evidence to support that providing transparent AI agents helped individuals gain health awareness, engage in conversations about healthy living practices, and promote collaboration among family members. Our findings provide insights into developing effective family-centered health interventions that aid non-collocated families in cultivating health together. The experiment results help to explain how transparent AI agents could mediate the health conversation and collaboration within non-collocated families.

4 Usability, Explaniability and Causability

The two case studies present our preliminary findings to support our arguments on the XAI is more than a matter of accurate explainable or interpretable models. Here we would like to draw a historical analogy to usability. One tension

in contemporary AI is the perception that core system qualities like speed, efficiency, accuracy and reliability might be compromised by pursuing objectives like transparency and accountability for some form of diffuse *explanatory value* [9]. But though our understanding of qualities like transparency and accountability is limited, this can be directly addressed to enhance the causability.

The trend of Explainable AI can be seen as analogous to usability: merely simplifying a user interface (in a logical/formal sense) may or may not make it more usable, instead the key to usability is a set of pragmatic conditions. It must be satisfying, challenging, informative, intuitive, etc. We could conclude that XAI is more than a matter of accurate and complete explanation, that it requires pragmatics of explanation in order to address the issues it seeks to address. One specific issue in XAI is that AI should be able to explain how it is fair. Such an explanation will necessarily intersect with an accurate system model but would be much more focused on interaction scenarios and user experiences.

On the history in age of 1980 simple noting of usability. Only saying keep simple as stupid? Directly pursue the simple solution is not the same as usability. User's ability to trust of understand AI is not sufficient. Usability we don't really have a theory in these aspects. Usability is not equal to empirical evidence, to do experiment with kind of explanations and exploratory interaction and explore the consequence. The consequence could be part of the usability. Something goes wrong, and the users need an explanation, i.e., we want to know what is happening. Explanations could be an engagement. Active thinking and active learning, user interaction and usability, and wrong and addition situations. Try to understand the system model, but why do uses want to get this explanation?

Carroll and Aaronson [3] investigated a Wizard of Oz simulation of intelligent help. They studied interactions with a popular database application and identified 13 critical user errors, including the application state people were in when they made these errors. In this way, the help simulation recognized and guided recovery from a set of serious mistakes. Carroll and Aaronson designed two kinds of helpful information: "how-it-works," explaining how the system model worked to allow the error and leaving it to the user to discover what to do, and "how to do it," describing procedures the user should follow to recover from the error and continue their task. They found that people often preferred help messages explaining how the database application worked, for example, when it noted the distinction between forms and data when users entered both field labels and numeric values. When puzzled by the system, such interactions were satisfying to users, but "how-it-works" messages particularly pleased users in answering questions just as they were being formulated. Simplifying a user interface in a logical/formal sense may or may not make it more usable. Key usability also considers pragmatic conditions - systems must be satisfying, challenging, informative, intuitive, etc.

The field of human-computer interaction (HCI) coalesced around the concept of usability in the early 1980s, but not because the idea was already defined clearly, or could be predictably achieved in system design. It was instead because the diffuse and emerging concept of usability evoked a considerable amount of

productive inquiry into the nature and consequences of usability, fundamentally changing how technology developers, users, and everyone thought about what using a computer could and should be [4]. Suppose that AI technologies were correctly reconceptualized, including the capability to effectively explain what they are doing, how they are doing it, and what courses of action they are considering. Adequate, in this context, would mean codifying and reporting on plans and activities in a way that is intelligible to humans. The standard would not be a superficial Turing-style simulacrum but a depth-oriented investigation of human-computer interaction to fundamentally advance our understanding of accountability and transparency. We have already seen how such a program of inquiry can transform computing.

References

1. Amershi, S., et al.: Guidelines for human-ai interaction. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, p. 3. ACM (2019)
2. Anderson, A., et al.: Mental models of mere mortals with explanations of reinforcement learning. ACM Trans. Interact. Intell. Syst. (TiiS) **10**(2), 1–37 (2020)
3. Carroll, J., Aaronson, A.: Learning by doing with simulated intelligent help. Commun. ACM **31**(9), 1064–1079 (1988)
4. Carroll, J.M.: Beyond fun. Interactions **11**(5), 38–40 (2004)
5. Craik, K.J.W.: The Nature of Explanation, vol. 445. CUP Archive, Cambridge (1952)
6. Dragoni, M., Donadello, I., Eccher, C.: Explainable AI meets persuasiveness: translating reasoning results into behavioral change advice. Artif. Intell. Med. **105**, 101840 (2020)
7. Eiband, M., Schneider, H., Bilandzic, M., Fazekas-Con, J., Haug, M., Hussmann, H.: Bringing transparency design into practice. In: 23rd International Conference on Intelligent User Interfaces, pp. 211–223. ACM (2018)
8. Guy, I.: Social recommender systems. In: Ricci, F., Rokach, L., Shapira, B. (eds.) Recommender Systems Handbook, pp. 511–543. Springer, Boston, MA (2015). https://doi.org/10.1007/978-1-4899-7637-6_15
9. Hancock, J.T., Naaman, M., Levy, K.: Ai-mediated communication: definition, research agenda, and ethical considerations. J. Comput. Mediat. Commun. **25**(1), 89–100 (2020)
10. Herlocker, J.L., Konstan, J.A., Riedl, J.: Explaining collaborative filtering recommendations. In: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work, pp. 241–250. ACM (2000)
11. Herring, S.C.: Computer-mediated communication on the internet. Ann. Rev. Inf. Sci. Technol. **36**(1), 109–168 (2002)
12. Hilton, D.J.: Conversational processes and causal explanation. Psychol. Bull. **107**(1), 65 (1990)
13. Holzinger, A., Carrington, A., Müller, H.: Measuring the quality of explanations: the system causability scale (scs). KI-Künstliche Intelligenz **34**(2), 193–198 (2020)
14. Knijnenburg, B.P., Bostandjiev, S., O'Donovan, J., Kobsa, A.: Inspectability and control in social recommenders. In: Proceedings of the Sixth ACM Conference on Recommender Systems, pp. 43–50. ACM (2012)
15. Levy, K., Barocas, S.: Designing against discrimination in online markets. Berkeley Technol. Law J. **32**(3), 1183–1238 (2017)

16. Liao, Q.V., Gruen, D., Miller, S.: Questioning the AI: informing design practices for explainable AI user experiences. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–15 (2020)
17. Liao, Q.V., et al.: All work and no play? In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1–13 (2018)
18. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. **267**, 1–38 (2019)
19. Mittelstadt, B., Russell, C., Wachter, S.: Explaining explanations in AI. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 279–288 (2019)
20. Ngo, T., Kunkel, J., Ziegler, J.: Exploring mental models for transparent and controllable recommender systems: A qualitative study. In: Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, pp. 183–191 (2020)
21. Norman, D.A.: Some observations on mental models. Ment. Models **7**(112), 7–14 (1983)
22. O'Donovan, J., Smyth, B., Gretarsson, B., Bostandjiev, S., Höllerer, T.: Peerchooser: visual interactive recommendation. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1085–1088. ACM (2008)
23. Powley, L., McIlroy, G., Simons, G., Raza, K.: Are online symptoms checkers useful for patients with inflammatory arthritis? BMC Musculoskelet. Disord. **17**(1), 362 (2016)
24. Ruben, D.H.: Explaining Explanation. Routledge, London (2015)
25. Tintarev, N., Masthoff, J.: Explaining recommendations: design and evaluation. In: Ricci, F., Rokach, L., Shapira, B. (eds.) Recommender Systems Handbook, pp. 353–382. Springer, Boston, MA (2015). https://doi.org/10.1007/978-1-4899-7637-6_10
26. Tsai, C.-H., Brusilovsky, P.: The effects of controllability and explainability in a social recommender system. User Model. User-Adapt. Interact. **31**(3), 591–627 (2020)
27. Tsai, C.H., You, Y., Gui, X., Kou, Y., Carroll, J.M.: Exploring and promoting diagnostic transparency and explainability in online symptom checkers. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–17 (2021)