ISSN 2186-7437

NII Shonan Meeting Report

No. 200

Social Explainable AI: Designing multimodal and interactive communication to tailor human–AI collaborations

Kary Främling
Brian Y. Lim
Katharina J. Rohlfing

September 18–21, 2023

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-Ku, Tokyo, Japan

Social Explainable AI: Designing multimodal and interactive communication to tailor human–AI collaborations

Organizers:
Kary Främling
Department of Computing Science, Umeå Universitet, Sweden
Brian Y. Lim
National University of Singapore, Singapore
Katharina J. Rohlfing
Transregional Research Center *Constructing Explainability*,
Paderborn University, Germany

September 18–21, 2023

Abstract

In recent years, research on explainable AI (henceforth, XAI) has intensified, responding to the societal challenge that many algorithmic approaches (such as machine learning or autonomous intelligent systems) are rapidly increasing in complexity, making it difficult for citizens to understand their assistance and to accept the decisions they suggest. There now exists a large body of approaches pushing forward many ideas of how algorithms should be made explainable or even be able to explain their own output. However, the users' perspective is rarely taken into account seriously. Obviously, not only research dimensions but also putting into operation a true involvement of the users are missing.

In this meeting, we therefore gathered scholars from different disciplines to account for the question of how explanation generation can be tailored to the users and their way of understanding. Starting this 4-days meeting, we proposed that, rather than being 'delivered' by the explainer, explanations become tailored when they emerge at the interface between the explainer (or explaining system) on the one hand and explainee (the receiver of an explanation) on the other hand. Both are active participants in shaping explanations during a social interaction. We thus considered social interaction to be the key to the involvement of the users. Derived from the state of the art, we structured our meeting along three facts concerning social interaction in general and explanatory dialogues in particular that became the structuring dimensions of our approach:

- *Interaction is multimodal* (e.g., visual, verbal, auditory), so XAI needs to account for different modalities of communications that are used in the process of constructing an explanation.

- *Interaction is incremental* and builds on the contribution of the involved partners who adapt to each other. In this sense, an explainee can also contribute to a successful explanation, e.g., by asking questions or by providing feedback regarding their understanding.
- *Interaction is patterned* in the sense that different contexts and goals will lead to the emergence of different social roles impacting the construction of explanations. In the case of an explanation, specific conversational patterns are followed, e.g., patterns of explicating the relations.

We achieved to agree on a conception of **social XAI** that is based on research in social interaction, and we raised awareness for this perspective among the participating scholars. We worked out a structure of a handbook for social XAI that will be published to keep the emerging community engaged. With the handbook, it is our objective to extend current research in computer science and offer new answers to the abovementioned societal challenge.

Background and Introduction

In our digitized society, computational approaches (such as machine learning or autonomous intelligent systems) are rapidly increasing in complexity, making it difficult for citizens to understand their assistance and to accept the decisions they suggest. In response to this societal challenge, research on explainable AI (or XAI) has intensified, pushing forward many ideas of how algorithms should be explainable or even be able to explain their own output. Consequently, recent work on XAI has broadened its perspective, tackling topics such as verbal explanations, interactivity of systems, hybrid approaches combining reasoning and learning for XAI, and the relevance of explanations to the users.

Whereas the emerging XAI approaches are concerned with interpretability or explainability, more recently, state-of-the-art research reveals a lack of context-awareness [1], a lack of interaction as well as personalization as reasons for why an explainable system is of little use to the users [2]. Responding to this gap in more details, current reviews define dimensions of which to consider the users' perspective, but also identify a gap of approaches toward the involvement of users. Clearly, an explanation addressing the *why* does not necessarily lead to an understanding. Instead, the interaction of current XAI systems is severely limited because they can only deliver an explanation, without tailoring it to the receivers' understanding and informational needs, nor to the given context. Responding to this limitation, Miller [3] argues that XAI can benefit from social science research. In this line of research, social interaction seems to be the key for making an explanation provided by a system understandable and relevant. The big advantage of viewing social XAI as *socially interactive explainable AI* is that concrete operationalizations for how to involve the users can be derived from research on social interaction. This is indicated in [4] proposing a framework for a social design of XAI. For this, the properties of an interaction need to be considered. These properties go beyond what has been suggested by Miller viewing explanations as selective and contrastive to foster the cognitive process of abductive reasoning (i.e., deriving a hypothesis to explain observed phenomenon). In fact, Brasse and colleagues detect that mainly cognitive theories are employed in XAI [5] that limit the current perspective. To solidify theoretical foundations, a broader and interaction-oriented perspective is needed.

To structure our view along clear dimensions and add to characteristics of an explanation proposed before, we focused on the following properties of interaction:

- **multimodality** (e.g., visual, verbal, auditory), so XAI needs to account for different modalities of communications that are used in the process of constructing an explanation.
- **incrementality**, which means that an interaction consists of elements that follow on each other allowing to build on the contribution of the involved partners who adapt to each other. In this sense, an explainee can also contribute to a successful explanation (e.g., by asking questions or by providing feedback regarding their understanding).
- **patternedness** in the sense that different contexts and goals will lead to the emergence of different social roles impacting the construction of explanations. In the case of an explanation, specific conversational patterns

are followed that also give rise to cognitive patterns (e.g., relations about the observed phenomenon will be explicated yielding cognitive operations that are typical for explanations and abductive reasoning).

Whereas in our discussions during the meeting, we agreed on a conception of social XAI that is based on research in social interaction and raised the awareness for this perspective within the participating scholars, our aim was to go beyond our meeting and to foster the emerging community in a sustainable way. Therefore, we worked out a structure of a handbook for social XAI that will be published to keep the emerging community engaged. With the handbook, it is our objective to extend current research in computer science and offer new answers to the abovementioned societal challenge by contributing to the development of:

- a multidisciplinary understanding of the mechanisms involved in the process of explaining, tailoring it to the process of understanding,
- computational models and complex AI systems that focus efficiently on what kind of explanation a person requires in a current context, and by
- multimodal interactions to achieve an interaction that unfolds over time and that makes a joint construction of relevant explanation patterns possible.

Overview of Talks

What's New in Social Aspects of XAI?—Current Research Gaps

Katharina J. Rohlfing, Transregional Research Center *Constructing Explainability*, Paderborn University

For our meeting, we started with the definition of artificial intelligence (AI) as intelligent systems that interpret data, learn from it, and use this learning to achieve specific goals and tasks [6]. According to the literature (e.g., [7]), weak and strong AI systems can be differentiated: Whereas weak AI is developed for specific tasks, strong AI is considered to be more flexible. The black-box nature of these models or their complexity is opaque not only to users but also to their developers. The models bear risks of biases and deskilling [7] that are based on system's (e.g., based on correlations) or human's behavior (e.g., overreliance).

The XAI is viewed as the solution to these problems. XAI is referring to "many user-centered, innovative algorithm visualizations, interfaces and toolkits" that support users with various levels of AI literacy to enable them to understand and trust [8, p. 1]. In fact, explainability is considered a prerequisite for fair, accountable, and trustworthy AI, eventually affecting how we manage, use, and interact with it [7]. XAI is applied in "diverse subject domains, from the bank customer who is refused a loan, the doctor making a diagnosis with a decision aid, to the patient who learns that he may have skin cancer from a smartphone photograph of his mole" [8].

Whereas XAI is viewed as solution to the problem of opacity, it has to date been mostly investigated with a method-oriented focus for developers in computer science [7]. In his seminal work, summarizing existing research on XAI, Miller [3] called for explanations that have to become more social in order to be relevant and helpful. Since then, research has turned to this objective attempting to address this call.

In this introductory talk, a summary of six recent overview articles was given to the participants. This way, a ground was provided for discussions about the specification of the term "social XAI" and what social aspects are needed in XAI.

- The paper by Meske and colleagues [7] presents the process of decision-making in which an XAI is involved. The innovation of the paper is to emphasize the **management of organizational processes** as a goal of XAI. It is based on the argument that XAI as a new form of material agency in organizational processes changes work routines. Groups of **stakeholders** are identified that should be regarded for the new goal of the XAI (Figure 3 on p. 58): AI regulators, AI developers, AI manager, AI users, and individuals affected.

- From the beginning, the paper by Ali and colleagues [9] differentiates between explainability of either the data, model, or post-hoc statements (Table 2 on p. 10). With this differentiated view, it provides an extensive overview on the variety of methods that are available in XAI criticizing that many methods already exist and can elicit explanations of various kinds. However, the methods achieve it without regard to whether they satisfy the intended audience's needs [9] as can be seen by the authors' analysis of

the methods presented in dependence on the users and their satisfaction. This way of presentation is derived from the original goal of XAI research which is to make AI systems more comprehensible and transparent to humans without sacrificing performance. The goal is reached when humans understand and trust the AI solutions. In Figure 27 (p. 32), the authors summarize the assessment methods currently available. The purpose of this extensive presentation is to define **how to attain trustworthy AI**. In Figure 29 on page 39, this paper also provides a landscape of the decision-making process that XAI supports. This landscape informs about what groups of users need to be taken into account. Whereas this paper manages to propose that trustworthiness is about explainability and how to assess it within a whole decision-making process, it does not go beyond pointing out that the receiver of an explanation has to be taken into account more.

- The paper by Brasse and colleagues [5] provides an **analysis of the application scenarios** that are considered in current XAI research. It successes in communicating to what extent different types of users (developers, experts, or lay users) are taken into account and in what kind of scenarios. Interestingly, the domains of healthcare and finance seem to be predominant in most of the XAI-studies. From their analysis, the authors conclude that XAI is "not very theory-rich" [5, p. 25] and mainly cognitive theories are employed. Clearly, there is a gap on social aspects of explanations and their long-term influence.

- The title of the paper by Haque and colleagues [10] promises to take the user's perspective. Indeed, the paper is about understandability and it departs from the XAI goal to empower users to adopt AI-based systems. The authors focus on **users' mental models including their beliefs and perception about the external world**. It suggests that users can be grouped accordingly when developing an XAI for specific purposes. This approach, however, results in a static XAI that is designed for a target group but is not able to fine-tune or modify its predefined actions during an interaction. In their suggestions for future research, the authors raise interesting novel questions such as what is the impact of AI on low-literate persons.

- In their paper, Arrieta and colleagues [11] provide concepts and taxonomies that guide toward responsible AI with the goal to place **audience as a key aspect** to be considered when explaining a machine learning model. This way, they address the research gap to involve the user's perceptive. For their purpose, in Figure 2 on p. 83, they not only identify different groups of the target audience (domain experts, regulatory entities, managers, data scientists, developers, product owner, and users affected by model's decision) but also define different levels of transparency, on which basis they develop their principles of responsible AI. Accordingly, a responsible AI must not discriminate (fair AI), take audience into account (transparent AI), generate benefit (human-centric AI) and consider privacy and security. This focus on responsible AI is clearly advancing current approaches but the question of how the audience can be taken into account remains unanswered.

- Providing some concrete answers to the question of how the audience or a user can be taken into account, the paper by Chromik and Butz [12] offers design principles for explanation user interfaces. Specifically, it takes **a systematic look at the way the interaction is designed in the XAI research**. For their purpose, the authors define explanation user interface (XUI) as "the sum of outputs of an XAI system that the user can directly interact with". With this definition, the explanatory power is assigned to the interaction and not to a statement. The paper introduces a taxonomy of interaction, according to which the applied XUI can be categorized. The taxonomy corresponds to metaphors with which interaction can be viewed as: information transmission, dialogue, control, experience, behavior optimization, tool use or embodied action (see Table 2 on p. 14 for an overview). The different categories of the interaction differ with respect of the system's interactivity, i.e., how much the user can be involved in the interaction and influence the system's output. The authors propose that for a system to be interactive, it needs to take advantage of the naturalness of a human–human dialogue, its incrementality allowing follow-up on initial explanations, flexibility through multiple ways to explain, and sensitivity to the user's mind and context of explanation. The gain of the paper lies in concrete suggestions on how a user's involvement can be achieved and how to operationalize it. It reveals, however, that there exist little approaches allowing preemptive task co-management and shared progress tracking in a human–XAI interaction, which abilities make a system flexible and allow an explanation to be tailored to the user.

In summary and following the conclusions by Ali and colleagues [9], most research on how to interpret and explain AI systems is mainly motivated by requirements of the developers rather than users. However, to comprehend an AI system satisfactorily, each user needs a different level of explanation [2]. For these levels, many solutions were proposed [9, 10, 11], however they rather categorize the users as a group that has particular characteristics and needs that require a specific way of explaining. More flexible systems are still lacking and barely respond to the original call for explanations in which the explainer and explainee interact with each other [3]. This vivid process, in which both partners contribute to the explaining, was recently described in details in natural everyday explanations [13].

It can thus be concluded that whereas social aspects are well-recognized in current XAI research and partially identified, what is missing is a stronger guidance toward how to involve users in a meaningful social interaction in order to make an explanation relevant and tailored to the user's emerging understanding.

For this meeting, it is proposed that what is lacking are the system's abilities to be multimodal, incremental, and patterned. These necessary properties were further described in the invited talks below.

Multimodal Interaction Shapes Explanation Dialogues

Angela Grimminger, Paderborn University
Hendrik Buschmeier, Bielefeld University

This talk addressed the multimodality of human interactions and how this is important for the co-construction of dialogical explanations. In this talk,

multimodality was defined as the different forms of verbal and non-verbal signals that interlocutors use during communicative interactions. Social interaction naturally takes place 'face to face' and interlocutors produce and perceive a continuous stream of meaningful bodily behaviour [14]. The modalities used in social interactions can be vocal/aural (including speech and prosodic features) or visuospatial (including hand gestures, head gestures, facial expressions, body postures, actions and gaze) [15, 16]. All of these behaviours in human interaction can be produced or perceived, for example, as signals of attention [17], for turn management [18], or as feedback [19, 20] and thus shape the ongoing interactions. Various multimodal means and their contribution can already be observed in the earliest parent-child interactions [21, 22, 23]. Using a video example of a dialogical explanation of a board game, the talk showed how non-verbal and multimodal forms of feedback from an explainee signal both understanding and misunderstanding, and how these signals can influence the explainer's ongoing explanation.

Interaction Is Incremental

Kary Främling, Umeå University

The human brain is presumably the least well understood "black-box model" on Earth. Despite this, humans are capable of introspection and self-reflection that allows them to justify and explain their reasoning and decision-making in various ways, depending on the explainee, the context of the explanation, and other factors. Human explainability is a social interaction whose objective is often to convince the explainee about the rationality of the explainer's reasoning.

Current state-of-the-art eXplainable AI (XAI) methods lack such interactivity. Even though it might be possible to divide the explanation into smaller chunks, the mathematical limitations of the XAI methods may prevent them from taking into consideration interactions between features and to misleading explanations where feature influences cancel each other when combined into feature coalitions.

The Contextual Importance and Utility (CIU) method [24, 25] overcomes both of these challenges, as well as other challenges of current state-of-the-art XAI methods. The presentation shows how CIU can answer questions such as "Why?", "Why not?", "Why A and not B?" (contrastive) and "What if?" (counterfactual) using a human dialog for justifying the choice of a car as an example. CIU's "intermediate concepts" deal with feature interactions in coalitions of features in a mathematically consistent way. CIU also avoids the phenomenon of feature influences cancelling each other due to the use of classical definitions of "importance" and "utility" from Decision Theory, which are extended to deal with non-linear models such as those represented by neural networks and other machine learning methods [26].

Finally, the emphasis of the presentation is to advocate that XAI research should focus on what kind of questions humans tend to ask and how the understanding of the black-box model's reasoning can be built through an incremental interaction process. CIU provides building blocks to bridge the gap between such Social AI and XAI methods [27]. However, achieving a truly Social AI will require an extensive collaboration effort between different domains such as sociologists, psychologists, human-computer interaction specialists, cognitive scientists and computer scientists.

Interaction Is Patterned

Anna-Lisa Vollmer, Bielefeld University

In his qualitative studies of parent-child interactions, the US-American psychologist Jerome Bruner identified interaction patterns that he called *interaction formats* [28]. He identified game formats like peek-a-boo, and the nursery rhymes ride-a-cock-horse that follow a relatively fixed structure in interaction patterns, utterance choice, and prosody. He also observed such game patterns that emerged between child and caregiver and as such were individual for this dyad. Bruner found picture book-reading to also happen in a format. Formats have a variant surface structure that include the concrete observable utterances and behaviors of the interaction partners (Emma, look! with pointing | Looking| What's that? | Babble | Yes, a pineapple! | an apple | yes), and an invariant deep structure that pertains to the abstract interaction structure (attentional vocative | query | correct label | feedback). [29] have proposed the concept of pragmatic frames (PFs) for learning words. The PFs are a sequence of coordinated behaviors – actions and language – of both interaction partners. PFs are recurrent. They emerge over time and first occur in a specific context. They also link the surface behaviors to the underlying cognitive operations (locate and follow finger | segment and recognize object | recognize name | link object with name). The relatively fixed patterns, once established over repeated interactions, carry meaning and help the learner to pick up the only variable information – in the book-reading frame the object label – that the learner is supposed to learn. The roles in the frame can be swapped such that the caregiver can take over the role of the learner when they are not able to fully participate in the frame yet. In human–robot interaction, fixed interaction patterns or protocols are used in learning from human users. These are artificial as they are fixed and defined a priori by the developer, tailored to the underlying learning algorithm (cf. [30]). When collaborating on a joint task, common ground on familiar frames helps to negotiate the used frame. On a smaller time-scale, interaction patterns emerge when contingency is given over repeated interactions in collaborative activity in unfamiliar communicative situations that help to establish meaning and can be a measure of successful interaction [31]. In the discussion, the question arose whether the interaction pattern of the interaction between staff and client when going to the bakery was a frame or rather a script like the frequently mentioned restaurant script [32] as the typical sequence of events happening when going to the restaurant. Whereas scripts also involve non-communicative events and events are not constrained to interaction partners, but interaction partners might change going in and out of the interaction, PFs are specific to a pair (or group) of interactants and develop over time. In the restaurant script, roles are just like in the bakery pattern non-interchangeable.

List of Participants

- Rachid Alami, LAAS-CNRS, France
- Heike Buhl, Paderborn University, Germany
- Hendrik Buschmeier, Bielefeld University, Germany
- Kary Främling, Umeå University, Sweden
- Angela Grimminger, Paderborn University, Germany
- Joris Hulstijn, Université du Luxembourg, Luxembourg
- Sylvain Kubler, Université du Luxembourg, Luxembourg
- Amro Najjar, Université du Luxembourg, Luxembourg
- Katharina J. Rohlfing, Paderborn University, Germany
- Igor Tchappi, Université du Luxembourg, Luxembourg
- Kirsten Thommes, Paderborn University, Germany
- Anna-Lisa Vollmer, Bielefeld University, Germany
- Henning Wachsmuth, Leibniz University Hannover, Germany
- Britta Wrede, Bielefeld University, Germany

Meeting Schedule

Check-in Day: September 17 (Sunday)

- Welcome banquet

Day 1: September 18 (Monday)

- Introductory talk by Katharina Rohlfing on what's new in social aspects of XAI and current research gaps
- Introduction round with the question of "what is social XAI?"
- Invited talk by Angela Grimminger & Hendrik Buschmeier on "Multimodal Interaction Shapes Explanation Dialogues" and discussion
- Group activity using Miro Board: Find 5 keywords that characterize AI as multimodal!
- Group photo shooting
- Invited talk by Kary Främling on "Interaction is incremental" and discussion
- Group activity using Miro Board: Find 5 keywords that characterize AI as incremental!
- Invited talk by Anna-Lisa Vollmer on "Interaction is patterned" and discussion
- Group activity using Miro Board: Find 5 keywords that characterize AI as patterned!
- Joint look at the 3 dimensions and discussion

Day 2: September 19 (Tuesday)

- Discussion on the structuring dimensions
- Splitting into 3 groups that provide the dimensions of the social interaction within XAI: Multimodality, Incrementality, Patternedness
- Working in 3 groups: Structure of the handbook and potential entries within the dimensions
- Working in 3 groups: Writing 1 page of an executive summary of the dimension, agreeing on roles and responsibilities for follow-up

Day 3: September 20 (Wednesday)

- Discussion on the structure of the book, the publication format, and further contributors
- Working in 3 groups: Reviewing the one page of executive summary and writing a further version of it considering further contributors and missing aspects
- Excursion and main banquet

Day 4: September 21 (Thursday)

- Discussion on the content of the overarching sections such as Introduction and Evaluation
- Development of a plan for further actions
- Wrap up

Overview of the Planned Handbook of Social XAI

In the discussions, we agreed on the following target group, structure, and format of the book.

Target Group for the Handbook The book is for researchers with a background in Computer Science who are interested in social aspects of AI in general and XAI in specific. It should offer both a quick and a more elaborated access to the social aspects. Importantly, the readership should find concrete operationalizations in order to be able to regard the social aspects in their research. For researchers with a background in social sciences, this book should offer insights into the application of social aspects. It will highlight new facets of the aspects and new research questions that the application context reveals.

Structure of the Handbook The following structure is planned for the book. For more details on the chapters, see Section "Summary of New Findings":

1. Purpose of the handbook
2. Glossary
3. Introduction
4. Multimodality
   - 4.1. Related terminology
   - 4.2. Multimodality in agents
   - 4.3. Nonverbal signals
   - 4.4. Feedback and grounding
   - 4.5. Ambiguity of nonverbal signals
   - 4.6. Synchronization of nonverbal signals in production
   - 4.7. Theoretical aspects of multimodal processing
   - 4.8. Multimodality in explanatory interactions
   - 4.9. Visualization and the use of multimodality to explain
5. Incrementality
   - 5.1. Incremental communication
   - 5.2. Adaptation
   - 5.3. Model of explanandum
   - 5.4. Models of the interaction partner and situation
   - 5.5. History of interaction
   - 5.6. Generation of explanatory content and requirements for social XAI
   - 5.7. Exploration of explaining content
6. Patternedness
   - 6.1. Context for explanations
   - 6.2. Values and norms

6.3. Explainee's and explainer's roles and relationships

6.4. Explanation goals

6.5. Responsibilities and their dynamics in explanations

6.6. Engagement in explaining

6.7. Structures underlying explanations

6.8. Practices: How to establish an explaining practice

6.9. Risks in XAI

6.10. Chances of XAI

7. Evaluation

7.1. Measuring the quality of social XAI

7.2. Uncertainly

7.3. Ethical trade offs

In the introduction, the social XAI will be defined as a socially interactive XAI against the existing background and state of the art. We will argue that interaction is the key to involve user. We propose that, rather than being delivered by the explainer, explanations become tailored when they emerge at the interface between explainer and explainee, who are both active participants shaping explanations during a social interaction. Thus, in contrast to explanations that are implemented in current human-robot interaction (HRI) and human-computer interaction (HCI), we propose to view explanation as a process that has a direct influence on explanation generation.

With this argument, the focus on the social interaction within the context of explanations will be motivated for the book. This context will be identified as specific and contrasted with other dialogue forms. Two application real-life scenarios defining the application of XAI will be introduced to provide concrete examples that highlight the specifics of the context. The three guiding and structuring dimensions (multimodality, incrementality, and patternedness) will be introduced.

Finally, the intention of the format (quick access through short and extended version) will be explained and the target audience will be identified: For persons with technology background, the entries/chapters should offer clear operationalizations of the social aspects. For researchers with the background in social sciences, examples of how the context of XAI generates new questions will be made explicit.

Format of the Handbook The book follows a new format that allows for both, a quick (2 manuscript pages = 1000 words) and a more elaborated (8 manuscript pages = 4000 words) access to the social aspects of the XAI. Each entry will constitute a social aspect. Overall, the aspects will be structured along three dimensions (see above).

Summary of Discussions

The three guiding and structuring dimensions of interactive XAI systems – multimodality, incrementality, and patternedness – were discussed in groups critically.

Eventually, the participants of the meeting agreed on these dimensions, because having the function to cluster the Handbook's entries, they communicate the focus on interaction clearly. In addition, they capture the properties of a future XAI system well. More specifically, the property of an XAI system to take advantage of the multimodal signals a human user can send and receive was discussed as being essential for a successful social interaction as a constructive process. It should be highlighted that multimodality pertains to a context-dependent utilization of semantic information that can supplement, reinforce, or clarify what is already presented. Therefore, multimodal signals of communication need to be regarded when explanations are tailored to and will provide more relevance for the users.

The context dependency of multimodal signals is strongly linked to how information is unfolding (and therefore the dimension of incrementality). The property to interact in an incremental way is necessary for an XAI system to build on the contribution of the involved partners and to adapt to each other. As the interaction unfolds, it is likely that the sequence of actions is specific and follows a format of exchange (related to the dimension of patternedness). The interaction is patterned in the sense that different contexts and goals will lead to the emergence of different social roles impacting the construction of explanations.

Even though the three dimensions seem to be intertwined, each of them highlights specific social aspects.

Summary of New Findings

Multimodality

Angela Grimminger, Hendrik Buschmeier, Joris Hulstijn, Amro Najjar, Igor Tchappi

Social interaction naturally takes place in face-to-face communication (i.e., being physically co-present and sharing a referential space) where interaction partners are able to perceive each other through multiple sensory modalities. They can hear their partner speak, they can see their partner's face and body language. This means that communication uses multiple modalities, for example speech is accompanied by a manual gesture and prosodic marking. All are closely coordinated in their timing and work simultaneously. Explanations that people give each other (e.g., why there is a rainbow) make good use of these modalities (e.g., visualizing the angle at which sunlight falls on raindrops by drawing the trajectory in the air with a hand gesture). But it is not only the speaker who uses these multimodal signals. While both partners can see each other, listeners also use non-verbal modalities, for example to give feedback on their understanding (signaling positive understanding by nodding with their head) or to indicate that they would like to ask a question (by looking at the speaker and breathing in). To make XAI social, this part of the handbook argues that being able to use and perceive such multimodal signals when explaining AI behavior or decisions to humans is crucial and (most likely) beneficial, as it can enable explanations that are tailored to the individual interaction partners' and situational needs, state of understanding, emerging goals, etc..

- In Chapter M1, we will describe the terminology related to that part of the handbook. It introduces sensory-based modalities (vocal/aural, visuospatial), individual modalities to produce signals (face, gaze, hands, voice ...). It is described how they can be used unimodally, but crucially also how they can be combined to form multimodal signals in interaction.
- In Chapter M2, we will describe various artificial agents that might provide XAI-explanations, it could be disembodied agent (like a voice assistant or a chat bot) or embodied agents such as embodied virtual agents of social robots (co-located or remote). Embodied agents offer (and it might be argued necessitate) the possibility to communicate multimodally as well as to make use of space and position. Space and position are useful as cues for identifying roles (link to roles chapter). Embodiment also has disadvantages. It creates expectations that may not always be met. For example, noise of the robot movement may hinder speech recognition. Here we will discuss engineering limitations of the robotics platforms.
- In Chapter M3, we will describe the continuous nature of nonverbal signals in contrast with the often more discrete nature of verbal means of communication. [production–comprehension]
- In Chapter M4, we will write about multimodal feedback and grounding processes in explanatory interactions.
- In Chapter M5, we will focus on the ambiguity and inherent vagueness of nonverbal signals, challenges in processing them, but also the (deliberate) use of vagueness and ambiguity to further the communication. It could also consider incomplete explanation, lack of specificity, contradictory signals.
- In Chapter M6, we will describe how multiple modalities are orchestrated in production of behavior, specifically focusing on the synchronization between signals on different modalities and between interaction partners, the timing of signals as well as how they create and fit into the overall rhythm of the interaction.
- In Chapter M7, we will discuss theoretical aspects of multimodal processing including binding between modalities and other theoretical perspectives (Clark's layers, Levelt's dual coding theory, enactment, additive, compensatory, and integrated, multimodal-multimedia).
- In Chapter M8, we will inform the reader about how multimodality plays a crucial role in human explanatory interactions.
- In Chapter M9, we will discuss XAI needs for multi-modality. Often, XAI resorts to visualization and the use of multi-modality to explain the sophisticated decision of ML algorithms and the autonomous agents.

Incrementality

Kary Främling, Heike M. Buhl, Sylvain Kubler, Kirsten Thommes, Britta Wrede

Social XAI should operate on an incremental basis, mirroring the inherently gradual nature of human social interaction. When XAI engages in social exchanges,

it must be able to align with the step-by-step, iterative character of human communication. This incremental approach involves a dialogic process where explanations are shared in smaller portions, enabling a finely detailed and gradual co-construction of explanations between the user and the XAI system.

An incremental explanation strategy requires and allows both parties to adapt to each other and achieve increased understanding. An incremental explanation strategy also facilitates the identification of misunderstandings by directly monitoring each other's reactions, including verbal and nonverbal cues. When misunderstandings surface, the explainer can employ social techniques like repetition, correction or scaffolding, adding or changing the modality, changing the whole explanation strategy to accommodate the explainee's needs and expectations. This interactive process empowers the explainer to enhance the understanding of the explainee and update their model accordingly.

To accomplish this nuanced adaptation, the explainer necessitates not only a model of the subject matter but also an understanding of the explainee's current comprehension level regarding the subject. Successful adaptation also hinges on considering the broader context of the interaction, encompassing factors like social roles, physical surroundings, ongoing tasks, application risks (as defined by AI Act), and constraints faced by the explainee.

This contextual awareness extends beyond the immediate interaction and encompasses the history of interactions with the same explainee and other individuals within the same social network. As part of a social system, this demands consistency in the explanations provided to different explainees, ensuring not only trustworthiness but also averting potential controversies among social peers. For instance, explanations should not appear to be contradicting to two humans who belong to the same group and interact with each other.

A social XAI system should possess capabilities surpassing those of human-human interaction. It should deliver pertinent and accurate information, particularly information with legal implications, while also maintaining human agency, empowering individuals to request explanations and the option to have their data forgotten at any time.

Importantly, this process requires the underlying XAI system to also adapt to the explainee's needs by providing an interface to query different explanatory elements. Current XAI approaches mostly consist of one explanatory element like feature importance or pixel-wise relevance. However, the explanatory need of explainees may exceed this uni-dimensional approach in order to better grasp the complex causal or relational landscape of the problem domain at hand.

Over time, the explanation process can be automatically improved concerning the optimal incremental steps in terms of quantity and content. However, the design must also systematically vary social explanations and explore whether new processes of incremental steps may improve the process and avoid path-dependency. This is important, for instance, if the global understanding of explainees has evolved over time, social changes require new approaches or also when other modalities of explanations become available.

- Chapter I1: Incremental Communication: will address the problem of chunking incremental reception and feedback moves.
- Chapter I2: Adaptation will be presented on the basis of monitoring (social cues); the challenge of finding misalignment and misunderstanding and how to scaffold understanding will be presented

- Chapter I3: Model of the explanandum: will center around the explanandum (AI's, explainee's) and how is it dependent on the domain model and the situation
- Chapter I4: Model of the interaction partner and situation: will provide introduction into (fixed and dynamic) partner model(s) and how they change depending on the perspective taking as a cognitive means.
- Chapter I5: History of Interaction: will address both levels, micro and macrolevel, of repeating or unfolding interaction. It will be of question how consitency and stability can be established at the macro level. Concepts of trust, mistrust, trust calibration, and reliability will be discussed against the evolution of social relations between AI and explainee.
- Chapter I6: Generation of explanatory content and requirements for Social XAI: will center around the explanandum model and exemplify different kind of explanations in dependence of vocabulary, abstraction level, and modality
- Chapter I7: Exploration of explaining content: will address the possibility of dynamically exploring the content under the influence of feedback about the understanding.

Patternedness

Anna-Lisa Vollmer, Rachid Alami, Katharina J. Rohlfing, Henning Wachsmuth

Social interaction does not evolve randomly. It follows patterns that have been and are established over time and that are further evolved. The same is true for explaining process. To lay the ground of research towards social XAI systems, this part of the handbook is about the ingredients to shape and conduct interaction in explaining processes. These ingredients are necessary to form patterns that emerge and help partners to act according to their expectations and roles. Some of the patterns are general to the interaction of two or more people, while others relate to the specific contexts, goals, and practices of explaining.

The interaction is organized sequentially around a goal. On this goal, the partners need to agree and maintain it, all along the process of interaction. Thus, the concept of the goal is central to an interaction being organized in a pattern. A goal of the human might be to understand the function of the system or to be informed about uncertainty in the output of the system. The organization toward a goal happens in a social and physical context and depends on it. Whereas social context characterizes the explainee's beliefs, social roles, and expectations about the dialogue, the physical context relates to relevant aspects of the actual dialogue setting at hand, including the topic being discussed as well as the relevant physical environment in which the dialogue happens. Overall, the process of explanation is structured in joint sequential interaction patterns that are shaped by a certain background of norms and values, and social or situational context. Patterns emerge and evolve over time within one explanation as well as across explanations. On a broader scale, more universal patterns that are not specific to a set of interaction partners exist (for instance those that are culturally transmitted).

Certain patterns are bound to a certain context such that contextual information can be used to determine which patterns to follow. However, the organization toward a goal is providing clear responsibilities to the participants for its achievements. Because of the sequential structure of the interaction (actions being performed sequentially), the responsibilities can be distributed differently among participants. Thus, partners need to be engaged in the interaction and committed to the goal.

It is important to ensure that for the humans, the interactive process will offer the possibility to contribute to the goal but also latitude to express, all along the process, their preferences to determine their contributions. In XAI, responsibilities, contributions and preferences of the explainee are taken into account continuously. It is the duty of the XAI to assist and facilitate the human contribution all along the process and to ensure that explainee is committed, accepts and is willing to continue.

The operationalization of structures, practices, and conventionalizations in XAI systems entails risks. These include that XAI may even foster the deskilling of people—an effect that can be expected in case that AI systems take over more and more cognitive tasks as well as in cases in which too much trust exists in the alleged reliability of the systems' outputs. It is important to proactively work on preventive measures for these and other risks, both on the side of the developers of XAI methods and by the deployers of XAI systems in the real-world scenarios. Only then, the great potential that XAI systems exhibit can be leveraged. A successful realization of social XAI will enable people to solve their tasks more efficiently and effectively by making best use of the inherent capabilities of AI methods, in professional situations as well as in personal matters. At the same time, it allows bringing in the humans' genuine skills and leaving control over the process to them wherever needed. The understanding of the AI's decisions and behaviors gained through social XAI will give people the trust needed to responsibly and maturely integrate AI into the cognitive processes of everyday life, but it will also make them aware of the aspects in life that distinguish artificial intelligence from human intelligence.

The part argues that interaction patterns are a beneficial means in explaining that can be transferred to XAI. When conventionalized global and more fine-grained organization of explanations is assumed, explanations can be generated more easily in such a sequentially organized interaction and at the same time, multimodal cues and signals of the human interaction partner can be interpreted more easily.

The planned preliminary section organization is as follows. The text above is largely aligned with this organization:

- Chapter P1: Context. This chapter will introduce the notions of social and physical contexts: social context is formed from cultural context and is applied in a situation with physical constraints that constitute the physical context. It will be specified how an explanation is embedded into a specific context and how these contexts change depending on the scenario and application domain.

- Chapter P2: Values and Norms. We will point to the fact that social contexts are relying on values and norms that are established in the society or a group.

- Chapter P3: Roles and Relationships. We will define the notion of social roles and role relationships (doctor–nurse, doctor–patient) as well as interaction roles (speaker, addressee, overhearer), all at various levels (utterance, segment, session, relationship). Roles are part of a script or frame, which can be formalized as a dialogue game (see P7). Roles are played by specific agents, but only if they qualify. Given a role, it is clear what the responsibilities and capabilities are (see P5). E.g., the speaker is allowed to indicate who takes the next turn.

- Chapter P4: Goals. In this chapter, we will discuss how the goal of the explanation process has to be established and agreed upon by the human and the machine and how it will be maintained and eventually refined or updated depending on the course of the explanation process.

- Chapter P5: Responsibilities. This chapter builds on the known concept of *structures* and shows that while in an interaction, a structure is followed, it can be dynamically distributed to the actions of the partners. This requires a variability in the responsibility for the goal (e.g., a partner might take more responsibility scaffolding the other partner's action or performing a lot of actions necessary to achieve the joint goal)

- Chapter P6: Engagement. We will not only point out that an agent (a human or a system) needs to engage in an interaction in order to commit to the goal, this chapter will further link to the concept of *human agency* (see below under "notes") and how the human should stay the master of the game.

- Chapter P7: Structures. This chapter will present recurring structures of interactions (and their goals) as they occur in explaining processes, from the single dialogue acts and signals in individual turns and micro-patterns, to the pragmatic frames that constitute the composition of series of interaction mechanisms, to patterns of common types of dialogue as a whole. It will establish abstract patterns visible across all or certain contexts and will detail how they are instantiated for specific social and situational contexts given.

- Chapter P8: Practices and Conventionalizations. In this chapter, we will talk about the formation of interaction patterns and the role of repetition in this process. We will provide an account on how practices and conventionalization pertain to explaining and to identify important prerequisites or 'biases' for social XAI systems.

- Chapter P9: Risks. This chapter will highlight main risks arising for individuals and societies through the increasing prevalence of AI systems, focusing on those that are particularly relevant or even amplified in the context of XAI. It will shed light on what to particularly pay attention to in order to alleviate these risks as far as possible.

- Chapter P10: Chances. Finally, this chapter will elaborate on the chances that emerge from a systematic design and realization social XAI for the acceptance and effective use of AI in professional applications and society.

Evaluation

Kirsten Thommes, Joris Hulstijn, Henning Wachsmuth, Suzana Alpsancar

Whether or not social XAI is successful in terms of generating understandable explanation tailored to the user needs to be evaluated. In the best possible scenario, the explanation process would be self-optimizing in the long-run, consecutively improving the explanation procedure. Previous research measuring the interaction quality of systems frequently relies on correlations rather than on causal identification strategies. This is the case, for instance, if explainees are questioned about the perceived interaction quality, understanding or trust in the system post-hoc. Next to many commonly known effects of questionnaire studies, e.g. evaluator-demand effects, this approach also bears the problem of systematically missing causality and instead measures correlations only, resulting in a missed opportunity to improve the system. This chapter discusses processes of testing for causality instead of correlations. Moreover, we discuss the operationalization of potential outcomes of Social XAI, particularly explainees' attitudes such as trust, likeability of interaction, and aversion and behavioral responses such as understanding, reproducability of explanations, or decision-quality. Next to measures of effectiveness, the designer of Social XAI systems need to assess how they navigate ethical trade offs. For instance, one may ask whether Social XAI should mimic human social cues and lead explainees to anthropomorphize the system. On the one hand, human-like explanations may be more understandable than other types of explanations. On the other hand, anthropomorphize the system may eventually even result in false expectation such as empathy or emotions. We discuss the most common ethical trade offs and how to assess them.

- Chapter E1: Measuring the quality of social XAI. In this chapter, adequate measurements assessing the quality of social XAI will be identified and discussed. We will focus on tests for causality instead of correlations and typical evaluation criteria and their operationalization.
- Chapter E2: Uncertainty. In this chapter, uncertainty as a function of (X)AI is introduced and discussed.
- Chapter E3: Ethical trade-offs. We will tackle important ethical aspects such as accountability, transparency, inclusiveness of XAI systems. We will further explain some of the most common ethical trade offs in Social XAI and how to assess them.

References

[1] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling, "Explainable agents and robots: Results from a systematic literature review," in *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems* (N. Agmon, M. E. Taylor, E. Elkind, and M. Veloso, eds.), pp. 1078–1088, International Foundation for Autonomous Agents and MultiAgent Systems, 2019.

[2] K. Sokol and P. Flach, "One explanation does not fit all," *KI-Künstliche Intelligenz*, vol. 34, pp. 235–250, 2020.

[3] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[4] K. J. Rohlfing, P. Cimiano, I. Scharlau, T. Matzner, H. M. Buhl, H. Buschmeier, E. Esposito, A. Grimminger, B. Hammer, R. Hab-Umbach, I. Horwath, E. Hullermeier, F. Kern, S. Kopp, K. Thommes, A.-C. Ngonga Ngomo, C. Schulte, H. Wachsmuth, P. Wagner, and B. Wrede, "Explanation as a social practice: Toward a conceptual framework for the social design of AI systems," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 13, pp. 717–728, Sept. 2021.

[5] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodriguez, and F. Herrera, "Explainable artificial intelligence in information systems: A review of the status quo and future research directions," *Electronic Markets*, vol. 33, no. 1, p. 26, 2023.

[6] A. Kaplan and M. Haenlein, "Siri, siri, in my hand: Who's the fairest in the land? on the interpretations, illustrations, and implications of artificial intelligence," *Business Horizons*, vol. 62, no. 1, pp. 15–25, 2019.

[7] C. Meske, E. Bunde, J. Schneider, and M. Gersch, "Explainable artificial intelligence: objectives, stakeholders, and future research opportunities," *Information Systems Management*, vol. 39, no. 1, pp. 53–63, 2022.

[8] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim, "Designing theory-driven user-centric explainable AI," in *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–15, ACM, 2019.

[9] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Diaz-Rodriguez, and F. Herrera, "What we know and what is left to attain trustworthy artificial intelligence," *Information Fusion*, vol. 99, p. 101805, 2022.

[10] A. B. Haque, A. N. Islam, and P. Mikalef, "Explainable artificial intelligence in information systems: A review of the status quo and future research directions," *Technological Forecasting and Social Change*, vol. 186, pp. 122–120, 2023.

[11] A. B. Arrieta, N. Díaz-Rodriguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garciag, S. Gil-Lopeza, D. Molinag, R. Benjaminsh,

R. Chatilaf, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[12] M. Chromik and A. Butz, "Human-xai interaction: a review and design principles for explanation user interfaces," in *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3*, pp. 619–640, Springer International Publishing, 2021.

[13] J. B. Fisher, V. Lohmer, F. Kern, W. Barthlen, S. Gaus, and K. J. Rohlfing, "Exploring monological and dialogical phases in naturally occurring explanations," *KI - Künstliche Intelligenz*, vol. 36, pp. 317–326, Dec. 2022.

[14] S. C. Levinson, *Pragmatics*. Cambridge University Press, 1983.

[15] J. Holler and S. C. Levinson, "Multimodal language processing in human communication," *Trends in Cognitive Sciences*, vol. 23, no. 8, pp. 639–652, 2019.

[16] H. H. Clark and M. A. Krych, "Speaking while monitoring addressees for understanding," *Journal of Memory and Language*, vol. 50, pp. 62–81, 2004.

[17] E. A. Schegloff, "Body torque," *Social Research*, pp. 535–596, 1998.

[18] A. Kendon, "Some functions of gaze-direction in social interaction," *Acta Psychologica*, vol. 26, pp. 22–63, 1967.

[19] J. Allwood, S. Kopp, K. Grammer, E. Ahlsén, E. Oberzaucher, and M. Koppensteiner, "The analysis of embodied communicative feedback in multimodal corpora: A prerequisite for behaviour simulation," *Language Resources and Evaluation*, vol. 41, pp. 255–272, 2007.

[20] Z. Malisz, M. Włodarczak, H. Buschmeier, J. Skubisz, S. Kopp, and P. Wagner, "The alico corpus: Analysing the active listener," *Language Resources and Evaluation*, vol. 50, pp. 411–442, 2016.

[21] I. Nomikou, M. Koke, and K. J. Rohlfing, "Verbs in mothers' input to six-month-olds: synchrony between presentation, meaning, and actions is related to later verb acquisition," *Brain Sciences*, vol. 7, no. 5, p. 52, 2017.

[22] I. Nomikou, G. Leonardi, A. Radkowska, J. Rączaszek-Leonardi, and K. J. Rohlfing, "Taking up an active role: Emerging participation in early mother–infant interaction during peekaboo routines," *Frontiers in Psychology*, vol. 8, p. 1656, 2017.

[23] A. Grimminger and K. J. Rohlfing, "Entstehung multimodaler Sprachlehrstrategien in spezifischen Interaktionen," in *Lernen durch Vorlesen* (E. Gressnich, C. Müller, and L. Stark, eds.), pp. 94–109, Tübingen: Narr Francke Attempo Verlag, 2015.

[24] K. Främling, *Modélisation et apprentissage des préférences par réseaux de neurones pour l'aide à la décision multicritère*. Phd thesis, INSA de Lyon, Mar. 1996.

[25] K. Främling, "Explaining results of neural networks by contextual importance and utility," in *Rules and networks: Proceedings of the Rule Extraction from Trained Artificial Neural Networks Workshop, AISB'96 conference* (R. Andrews and J. Diederich, eds.), (Brighton, UK), 1-2 April 1996.

[26] K. Främling, "Decision theory meets explainable AI," in *Explainable, Transparent Autonomous Agents and Multi-Agent Systems* (D. Calvaresi, A. Najjar, M. Winikoff, and K. Främling, eds.), (Cham), pp. 57–74, Springer International Publishing, 2020.

[27] K. Främling, "Counterfactual, contrastive, and hierarchical explanations with contextual importance and utility," in *Explainable, Transparent Autonomous Agents and Multi-Agent Systems* (D. Calvaresi, A. Najjar, A. Omicini, R. Aydogan, R. Carli, G. Ciatto, Y. Mualla, and K. Främling, eds.), (Berlin, Heidelberg), pp. 180–184, Springer-Verlag, 2023.

[28] J. Bruner, "Child's talk: Learning to use language," *Child Language Teaching and Therapy*, vol. 1, no. 1, pp. 111–114, 1985.

[29] K. J. Rohlfing, B. Wrede, A.-L. Vollmer, and P.-Y. Oudeyer, "An alternative to mapping a word onto a concept in language acquisition: Pragmatic frames," *Frontiers in Psychology*, vol. 7, p. 470, 2016.

[30] A.-L. Vollmer, B. Wrede, K. J. Rohlfing, and P.-Y. Oudeyer, "Pragmatic frames for teaching and learning in human–robot interaction: Review and challenges," *Frontiers in Neurorobotics*, vol. 10, p. 10, 2016.

[31] A. L. Vollmer, J. Grizou, M. Lopes, K. J. Rohlfing, and P.-Y. Oudeyer, "Studying the co-construction of interaction protocols in collaborative tasks with humans," in *Proceedings of the 4th International Conference on Development and Learning and on Epigenetic Robotics*, pp. 208–215, IEEE, 2014.

[32] R. Abelson and R. C. Schank, "Scripts, plans, and knowledge," in *Thinking: Readings in Cognitive Science* (P. N. Johnson-Laird and P. C. Wason, eds.), pp. 151–157, Cambridge University Press, 1977.