Components of an Explanation for Co-constructive sXAI

Anna-Lisa Vollmer, Heike Buhl, Rachid Alami, Angela Grimminger, Axel-Cyrille Ngonga Ngomo

Chapter 3
Components of an Explanation for Co-constructive sXAI

Anna-Lisa Vollmer, Heike M. Buhl, Rachid Alami, Angela Grimminger, and Axel-Cyrille Ngonga Ngomo

Abstract The chapter "Components of an Explanation for Co-Constructive sXAI" examines the fundamental components that constitute explanations within the framework of social explainable AI (sXAI). It defines key concepts such as the explanandum (the entity being explained), the explanans (the manner of explanation), the explainer (the provider), and the explainee (the recipient), and it explores their interactions. The chapter emphasizes the complexity of explanations, highlighting the dynamic nature of the explainee's evolving understanding along with the contextual factors affecting the explanation process. It advocates an approach to co-constructed explanations in which the explanandum and the explanans adapt to the explainee's needs, allowing roles to interchange. This contrasts with traditional XAI methods that assume a static, one-way knowledge transfer. By focusing on the conceptualization of co-constructive explanation, the chapter aims to inspire more effective and human-centered AI systems, setting the stage for future research in and the following chapters on social XAI.

A.-L. Vollmer (✉)
Interactive Robotics in Medicine and Care, Medical School OWL, Bielefeld University, Bielefeld, Germany
e-mail: anna-lisa.vollmer@uni-bielefeld.de

H. M. Buhl
Institute for Human Sciences – Psychology, Faculty of Arts and Humanities, Paderborn University, Paderborn, Germany
e-mail: heike.buhl@uni-paderborn.de

R. Alami
Laboratory for Analysis and Architecture of Systems, Artificial and Natural Intelligence Toulouse Institute, Université de Toulouse, Toulouse, France
e-mail: rachid.alami@laas.fr

A. Grimminger
Psycholinguistics, Faculty of Arts and Humanities, Paderborn University, Paderborn, Germany
e-mail: angela.grimminger@uni-paderborn.de

A.-C. Ngonga Ngomo
Data Science, Heinz Nixdorf Institut, Paderborn University, Paderborn, Germany
e-mail: axel.ngonga@uni-paderborn.de

K. J. Rohlfing et al. (eds.), *Social Explainable AI*,
https://doi.org/10.1007/978-981-96-5290-7_3

3.1 How Does This Chapter Relate to XAI?

When dealing with sXAI, it is necessary to look at the components of explanations, the entities involved, the partners/agents, and so forth. What is an explanation anyway? What is explained? What means are used to explain? And who explains to whom?

This chapter is designed specifically for readers with backgrounds in AI and technical disciplines. It aims to go beyond definitions and create a more in-depth basis for the central terms of the book. Important words and concepts are (see Fig. 3.1) the following:

- Explanandum: the entity (event, phenomenon) that is the object of an explanation.
- Explanans: the (verbal) way that an explanation can be expressed and co-constructed by both partners/agents.
- Explainer: a human or nonhuman agent in the role of producing an explanation in a social interaction.
- Explainee: a human or nonhuman agent in the role of the addressee of an explanation.

However, it is possible and necessary to go further and ask, for example, what is the subject of an explanation? In the scenario with Kary (cf. Sect. 2.3), is it the car? Its specific features? Kary's preferences or decision? The reasons behind this? Because the explanandum may change over the course of an explanation (see Sect. 3.3), such that we here introduce the term *global explanandum* for the whole explanation as a sequence of explananda serving as subcomponents for individual explanation steps. Moreover, the roles of explainer and explainee are also not clear and fixed in explanation interactions. Note that either interaction partner can have the responsibility to explain. The explainer can be one person at one moment and another at the next. Thus, it may be the case that the explainee explains their goals or contributes partial aspects of the explanation if they have understood something well. There are even situations in which the roles of explainer and explainee are not fixed at all, but both work together toward understanding. This is often the case, for example, in learning settings. Thus, a further aim of this chapter is to increase awareness of the complexity of the topic. We aim to present a problem outline. For certain aspects, we shall pose questions and present possible options without offering definitive solutions. Some unresolved aspects will form the basis for future research endeavors. However, many challenges are also central topics in other chapters of this book to which we can refer already.

In Table 3.1, we shall exemplify the above components of an explanation in the scenarios presented in Chap. 2.

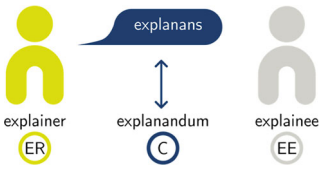

Fig. 3.1 Explanation terms

3.2 The Role of Explainer and Explainee

Figure 3.1 shows a minimalist explanation with one explainer (ER) and one explainee (EE). In human–human explanatory interactions, the explainers usually know more about a certain topic, the resulting explanandum, than the explainees (or they think they know more). The explainees strive for understanding and the explainers support them in overcoming a knowledge gap (cf. examples in Chap. 2). Regarding human–computer/robot interaction, in most cases, it holds that the explainer is the artificial agent that justifies its action, beliefs, decisions, and so forth. Thus, the characteristics of these roles go beyond the difference in knowledge. They are discussed in detail in Chap. 8. In classic XAI, the AI is the explainer, and the human user is the explainee. The AI explains or justifies its actions and decisions (see below regarding the explanandum). Further humans are often involved here as well. This is the case, for example, when an expert explains the AI's decisions to a layperson. However, it can be the other way around—for example, when a human explains to a robot why they do not like the recommendation given by the machine. Several further differentiations are important: There are differences between the characteristics of different explainers and explainees such as the type of the partner (robot, human, avatar, etc.), their domain knowledge, the explainers' ability to explain, their interest in the explanandum, and so forth (cf. Chap. 14). Besides the individuals, the relationships that explainer and explainee have with the explanation context differ and impact on the role of explainer and explainee (e.g., good friends, a student and their teacher, a doctor and a patient, a police officer and the suspect, etc.) (see Chaps. 4 and 8). The roles of explainer and explainee can even be reversed during an explanation. During some explanation steps, the original explainee might explain their point of view to the original explainer or might add some knowledge regarding the domain, the explanandum, or related topics. This leads to the need to define the explanation: Is it still the same explanation if the roles reverse or the explanandum changes (even slightly)? Additionally, explainers and explainees themselves change over the course of the explanatory dialog. The most obvious differences pertain to the explainee's level of understanding the explanandum—something that should increase over the course of an explanation. Likewise, there are changes regarding their cognitive load and attention as well as the resulting verbal and nonverbal behavior (see Chaps. 19, 13, Robrecht et al., 2024.) Another aspect to note is the number of people involved in the explanation: Fig. 3.1 shows two people involved. However, in many explanations—for example, at school or university lectures—one explainer and many explainees are the norm. This limits the possibilities for individualized adaptive explanations (cf. Chap. 13).

Table 3.1 The components of an explanation in the six scenarios presented in Chap. 2

Scenario	Explainer	Explainee(s)	Global Explanandum and Sequential Subcomponents
1	Kary	Marie, car-interested friend	Reason for buying Car A with sequential subcomponents: • Reason for buying Car A, more specifically restricted to what interests or is not known to or convinces the explainee • Reason for not buying Car B • Reason for not buying Car C • Reason for not buying Car C with extra option
2	Decision Support System	Kary	The current proposal of the AI system: • Reason for buying Car A • Reason for preference for Car B • Reason for buying Car C
3	Robot co-worker or service robot	Human co-worker or assisted user	Robot goals and plans, decisions and actions • This can be partially observed by the human explainee • Why and how robot has made a decision or achieved an action • Need for predictable, legible and acceptable behavior
4	Decision support system	Physician and patient	A diagnosis and treatment decision: • Reason for why proposing Diagnosis or Treatment A • Reason for why not proposing Diagnosis or Treatment B • Diagnosis or treatment for a different value of Feature X
5	RoboChef	User	Teaching a recipe to the user: • Why the user's dish is not like the robot's
6	Knowledgable human	Non-knowledgable human	Game explanation, consisting of subexplananda such as the following: • Goal of the game • Items such as cards, dice, meeple, objects, etc. • Rules and strategies

3.3 Explanandum

As stated before, the explanandum[1] is the entity (maybe an event, phenomenon, or decision) that is the object of an explanation (see Fig. 3.1, cf. Garfinkel (1981)). In scientific explanations, an *explanandum* is a term used to refer to the statement or phenomenon **that is to be explained**. It is the proposition, fact, or phenomenon that requires an explanation (Woodward, 1979). For example, if someone observes that an object falls to the ground when released, the explanandum in this case would be the falling of the object. Scientists then aim to provide an explanans, which is the set of statements or principles that constitute the explanation for the explanandum (Bokulich, 2011). In this case, the explanans might involve the principles of gravity and the laws of motion. In the following, we shall focus on everyday explanations instead of scientific ones. Everyday explanations do not claim to be complete and exact (Rohlfing et al., 2021; Miller, 2019). They often contain partial information. During the course of an explanation, the explanandum in everyday explanations is a "moving target" (Rohlfing et al., 2021, p. 720), not predefined and finished at the beginning of the explanation.

The explanandum has a global form describing what the explanation is about in general (*global explanandum*). However, we use the term explanandum to refer to a *local explanandum*, which is the content-related goal of one explanation *step*. We define an explanation step as one explainer move and the explainee's following contribution. The explanandum might be the same for multiple subsequent explanation steps. Over the course of an explanation, these local explananda form a sequence.

The explanandum is not only defined by a given situation,[2] event, or object (e.g., the car itself in Kary's car scenario Chap. 2) but also by the goals of the explainer and the explainee. For example, there are different forms of understanding, and it can be the goal of an explainee to *comprehend* something about cars or this specific car (to acquire knowledge) or to be *enabled* to drive cars or this car—that is, to learn an action (cf. Buschmeier et al., 2025). Regarding comprehension as well as enabledness, goals can differ widely with regard to the level of intended understanding, the specific object, and so forth. Also in this sense, the explanandum can change over the course of an explanation, and thus, it is co-constructed by explainer and explainee (cf. Chap. 13).

Different disciplines have different interpretations of how and – in a figurative sense – where this construction takes place. From the point of view of cognitive psychology, the explanandum can be seen as a mental representation of the explainer as well as the explainee (cf. Johnson-Laird, 2005). Thus, the explanandum is different for explainer and explainee. It can be expected that, at the beginning of an explanation, the explainer may have a better (more complete, more correct) representation than the explainee, whereas over the course of the explanation,

[1] The plural is explananda.

[2] The context significantly influences both the emergence and modification of the explanandum (see Chap. 4).

the representation of the explainee is built up (cf. Kintsch, 1998). Additionally, explainer and explainee might have their own models of what the explanation should be about and what the need for understanding in the explainee is. Following a constructivist view of knowledge and knowledge generation, an objective explanandum does not exist, just as the knowledge of two people is never the same. Similarly, computer sciences consider the explanandum as a mental model/domain model that is part of an explaining system (cf. Schmid & Wrede, 2022). Going beyond this individual constructivist perspective, a *social* constructivist view such as that in conversation analysis and sociocultural theory locates the explanandum in between the explainer and the explainee (Kern & Selting, 2020; Rogoff, 1998, e.g.,). As a co-construction, it emerges out of negotiation processes during the course of an explanation (cf. Rohlfing et al., 2021; Chaps. 1, 13).

Coming to explananda in sXAI, the sXAI system estimates the needs of the explainee.[3] Speaking about these 'needs,' we address different aspects such as the explainee's desires, goals, interests, knowledge gap, and so forth (Alpsancar et al., 2024). These aspects are built into the sXAI's model of the explainee (the *partner model*) and are used to derive the explanandum. These models of explainee and explanandum are reevaluated and adapted at each step of an explanation (see Chap. 14). Regarding the case of sXAI, the explanation need arises based on observations of AI outputs. Depending on the context and application, the explanation can take, in particular, these forms:

1. **Immaterial:** a decision that has been taken by the machine in response to a request of the user; a solution to a request formulated as a choice, a selection, a diagnosis, a plan, a recommendation or an assessment
2. **Concrete:** an object or a device built by the AI system (e.g., a 3D printed object)
3. **A past activity**: a task that has been achieved by an AI-enabled machine
4. **An ongoing activity:** the overall activity a machine is currently performing: its goal, the physical action the machine is currently performing (the current step in its plan) that can be observed together with its past actions, and its current plan for future action (including the anticipated contribution of the humans in case of a human-machine shared activity), with the latter presenting a fluid transition to the immaterial outputs

In the first three cases, the explanation process happens post hoc. In the ongoing activity case, the explanation activity accompanies a human–machine shared activity. It is somehow intermixed with it.

Importantly, the observed output might not be congruent with the respective global explanandum. Let us consider the example of multiple past actions of an AI-enabled machine as the output that the human explainee has witnessed. The

[3] We use the term "need" in a colloquial sense, because it is well-known and popular in XAI literature (e.g., recently Human & Watkins, 2023). However, it is problematic, because it seems like a scientific term used without a conceptual background and it implies the perspective of the user as a passive person.

explanation need arises, because the explainee thinks to have recognized some commonality in the past actions (i.e., an interpretation) and suspects a general strategy behind the machine's behavior. The global explanandum is, thus, the interpretation of the output (i.e., the general strategy) or even part of the machine's architecture causing the commonalities in the observed behavior.

The output is perceived, at least partially, by the human explainee as a physical object or as an action performed by the machine in the physical environment, on the human, or accompanying/supporting a human action using speech, text, or various modalities (see Chap. 23).

It is important to note that the explanandum might be perceived only partially (e.g., due to spatial conditions) and might need elaborate interpretation (cf. Chap. 4). This should be taken into account in the explanation process. Witnessing it partially might give rise to a fair amount of uncertainty, adding to the co-constructive nature of the explanandum.

It is also important to note that when the explanation accompanies an activity of the machine, the explanandum is somehow intertwined or even merged with the production of the explanans. The machine performs an action and conveys explanation information about it. An example is when a robot adapts a motion to achieve a task or synthesize it not only to achieve the task but also to convey to the human information about its intention or target.

Instances of Explanandum and Explanans for an AI-Enabled Robot

The following two cases illustrate a post hoc explanation process (Case 1) and an explanation of an ongoing activity (Case 2) (see Sect. 2.6 for a detailed example) for a cognitive and interactive robot (Lemaignan et al., 2017; Hellström & Bensch, 2018; Clodic et al., 2017; Arnold et al., 2021; Sakai & Nagai, 2022).

Case 1

Human H1 has given a task to Robot R. R has achieved the task.

1. This task involved R alone
2. This task involved assistance to Human H2 (for what concerns Human H2, refer to Case 2)

The task has been achieved. Some of its effects are perceivable by H1. H1 requests R to give an explanation about what it has done (what happened with respect to the task): The global explanandum is what happened—that is, what the robot has decided, what it has done, and what resulted in terms of effects in the environment and with respect to H2.

Case 2

Human H1 has given a goal to Robot R, and it is currently performing the task to achieve it

1. This task involved R alone but in the presence of a Human H2 (or H1 themselves)
2. This task involved a collaborative activity with H1

In this case, the explanans is closely intertwined with the explanandum (see Sect. 3.4). The task is ongoing. Some of its effects are perceivable by the human copresent with the robot (H2 and/or H1).

H1 requests R to give an explanation about what it is currently doing: The output is what happened until now. The global explanandum can be about the current activity and its goal and also about what the robot is planning to do.

Moreover, the robot should consider the effects of its behavior on copresent humans. At each step, while achieving its task, it has to ensure that its behavior is predictable, legible, and acceptable by copresent and coacting humans.

3.4 Explanans

As above with regard to explananda (cf., Sect. 3.3), for the sake of clarity, we also differentiate between global and local explanantia.[4] The global explanans addresses the (verbal) way that an explanation can be expressed and co-constructed by both partners (see Fig. 3.1). It is a sequence of local explanantia that determine **how** the respective explananda are conveyed at each explanation step. The way an explanation is expressed depends on a plethora of factors (see also Chap. 4). Environmental factors such as loud background noise influence how the explanation is given—in this case, the volume or other, nonverbal communicative means. Also, the norms applied in a setting – when explaining to a superior in a work context versus when explaining to a peer – impact on how an explanation is expressed, for example, which gestures and words are chosen. Explanations are expressed using multiple modalities: for example, with a graphic representation, pictogram, gestures, but also verbal utterances (cf. Chaps. 19 and 25). Due to the discretization of an explanation into explanation steps in our formalization (cf., Sect. 3.5), the production of the explanans is also discretized. However, it is produced continuously such that the explanans used is adapted online during the interaction based on observations of the explainee's reactions (Pitsch et al., 2009; Vollmer et al., 2014; cf. Chap. 13).

The explanans is closely intertwined with the explanandum. The explanandum was defined above as the entity that is the object of an explanation, whereas the explanans is the way this is done. Choosing and producing an adequate explanans might, however, involve referring to a different topic—for instance, by using a specific metaphor. This makes the distinction between explanans and explanandum difficult. In our view, and according to our definitions of the global explanandum and the local explananda, the explanans could involve a reference to a different topic within one explanation step. A broader use of a different topic that extends over multiple explanation steps and therefore is explained affects the respective explananda in that the novel topic will constitute new explananda. We argue that

[4] Explanantia is the plural of explanans.

the choice of the explanantia in the explainer is determined by an explanation strategy that couples explananda with the respective explanantia. The explanation strategy is responsible for not only current but also the planning of future explananda and explanantia, and it is updated constantly as the interaction unfolds. Further, explanations, being a form of social interaction, naturally take place in face-to-face communication (i.e., being physically copresent and sharing a referential space) in which (human) interaction partners are able to perceive each other through multiple sensory modalities and also use a variety of modalities to express themselves. Therefore, the global explanans, local explanantia, and the feedback from interlocutors (with respect to, e.g., the current level of understanding or the focus of attention) within the social interaction can be expressed verbally, nonverbally, or multimodally (with different combinations of multimodal behaviors) (see Chap. 19). Empirical research has provided evidence that multimodality facilitates cognitive processing and understanding (rather than hindering it, as one could expect because more information needs to be processed simultaneously) (see Chap. 18, e.g., regarding prosody Kern, 2007 or regarding gestures Holler & Bavelas, 2017).

Considering the factors influencing an explanans above, an explanans depends even more on the context than the explanandum.

For sXAI, the explanans plays a crucial role for conveying an explanandum and making explanations more human-centered. However, the (linguistic) dimensions and parameters of the explanans are still a subject of current research even in human interaction, making the development of a toolbox for different explanantia for sXAI especially difficult. In addition, multimodal communication becomes relevant for sXAI as soon as explanations are (the explanans is) produced by technical agents that provide anthropomorphic cues—for example, agents that use language, speech, or exhibit human features (behavioral and morphological factors; Kim & Im, 2023). Further, it is most likely that humans produce explanantia as multimodal utterances in which information might be distributed across different modalities also when interacting with technical agents. Therefore, a technical agent should be equipped with proper processing allowing it to interpret multimodal input and respond accordingly within explanatory interactions. However, given their ambiguity and their dependency on context and culture, the integration of multimodal aspects into technical systems is a challenging endeavor when it comes to both their perception and their production (cf. Chap. 19).

3.5 Operationalizations of the Explanation Components

This section aims to clarify the multifaceted nature of explanation by proposing a formal structure that captures its core components while also highlighting the intricacies that are frequently neglected in current XAI models.

To simplify the implementation of XAI systems, several strong assumptions are often introduced. These include idealized conditions such as perfect communication, fixed roles, and the absence of contextual factors. This contrasts sharply

with the incremental and context-dependent nature of human explanations. By outlining an exemplary set of these assumptions, we aim to illuminate not just how explanations can be formalized but also the complexity that lies beneath the surface—complexity that must be acknowledged and addressed for XAI systems to truly mirror the richness of human explanatory practices.

We can assume an explanation E to be a process that consists of a sequence of explanation steps $e_1 \dots e_n$. In each step e_i, the explainer er_i (i.e., ER for explanation step i) communicates the explanans es_i to the explainee ee_i (i.e, EE for explanation step i) according to an explanation strategy s_i. In each step, the explainee may also signal or express explanation needs (i.e., the explanandum), denoted em_i. Consequently, each step e_i can be regarded as a 5-tuple $e_i = (er_i, ee_i, s_i, es_i, em_i)$, in which an explainer provides an explanans to an explainee according to an explanation strategy, to which the explainee might form a need that could become apparent or identifiable through cues.

Several strong assumptions can simplify the job of explaining for XAI systems in contrast to the complexity of typical human–human explanations. Most importantly, we could assume the following:

1. The explainer er_i has the best interest of the explainee ee_i in mind and does not explain for the sake of explaining. Humans can explain for different reasons; see Chap. 2, but AI explains only to bring about understanding in the explainee (because of a need for understanding).
2. There are no errors in the communication of the explanandum—that is, the relation between er_i, ee_i, and em_i is functional.
3. There are no errors in understanding the explanans: es_i is understood exactly as intended.
4. er_i and ee_i are fixed roles during each step serving the knowledge transfer from er_i to ee_i.
5. There exists at least one optimal EM and the respective explananda em_i are fixed a priori and objective.
6. There is no error in monitoring the explainee (or no monitoring at all, when the ee_i have known changing states) and the next explanandum depends only on the current state of the explainee and is independent of the explanation history. The steps contain all information necessary to ensure $p(e_i|e_{i-1} \dots e_1) = p(e_i|e_{i-1})$.
7. No additional context factors have to be taken into account.

3.6 How Does This Chapter Inspire Further Directions of XAI?

In contrast to traditional state-of-the-art (SOTA) XAI approaches, which often rely on simplifying assumptions Sect. 3.5 to streamline the explanation process, this section delves into the complexities inherent in natural human–human explanations. Co-constructing social XAI challenges the above assumptions by emphasizing that

explanations are not static, one-way transfers of information, but rather evolving dialogs shaped by the mutual engagement of explainer and explainee. This section explores how abandoning assumptions such as fixed explananda (Assumption 5), rigid role assignments (Assumption 4), and static explanation steps (Assumption 6) opens up new possibilities for creating more adaptive, personalized, and context-sensitive XAI systems. By focusing on the alignment of beliefs and the negotiation of understanding, co-constructing sXAI aims to mirror the richness and complexity of human explanatory practices, pushing the field toward more human-centered AI solutions.

Some of the above assumptions in Sect. 3.5, especially 4, 5, and 6 breach the idea of co-constructing sXAI. In simple SOTA XAI, all of these assumptions and assumption 7 hold. Assumption 5 might often even be broader in SOTA XAI and pertain to an equality of all explainees such that the explanation produced is independent of the explainee. Personalized XAI does not assume equality of explainees and "[characterizes] an explainee as an individuum with preferences, personal characteristics, intentions, etc." (Rohlfing et al., 2021, p. 719). Still, explanations here are not given in a social interaction. In recipient-designed XAI (Miller, 2019), the above are not assumed directly. Miller (2019) supports the concept of recipient design by emphasizing the need to tailor AI explanations to the specific purposes and cognitive processes of human users. However, in contrast to co-constructing sXAI, the global explanandum is not a matter of co-construction, but the explanation is only an interactive 'transfer of knowledge,' and, as such, serves the successive filling of a knowledge gap. Accordingly, the following is assumed: EM and, thus, the explananda em_i depend on the current and successively updated model of the explainee, but the global explanandum is fixed a priori and objective.

In co-constructing sXAI, the global explanandum is not fixed over the course of an interaction, but a "moving target" (Rohlfing et al., 2021). As the explainee's understanding grows, new needs or wants for understanding may form. In some explanations, the explainee might initially not be sure what exactly they would like to know. In this case, the explanandum might emerge in the interaction between explainer and explainee. In other cases, the explanandum needs to be negotiated, because there might be an (initial) offset between what the explainer believes the explanandum to be and what the explainee believes it to be. Thus, in a co-construction, if assumption 5 does not hold, then the explanandum at each step i is not unique but rather consists of beliefs held by the explainer er_i and the explainee ee_i. Formally, assuming a belief function b, we get $em_i = (b(er_i), b(ee_i))$. Therewith, the signature of explanation steps is now $e_i = (er_i, ee_i, s_i, es_i, b)$.

A good explanation, then, is characterized by a good alignment of the respective explananda of explainer (i.e., $b(er_i)$) and explainee (i.e., $b(ee_i)$).

Additionally, in co-constructing sXAI, assumption 4 does not hold, and, for parts of an explanation, the explainee might become the explainer and vice versa, thereby reversing their roles (see Sect. 3.2). The explainee might, for example, become the explainer and explain their prior understanding or point of view to the explainer turned explainee.

Assumptions 2 and 3 may hold for co-constructive explanations. If they do not hold, however, the need for co-construction increases, because there are misunderstandings and miscommunications that make an alignment even more necessary.

These observations highlight the limitations of current SOTA approaches and suggest the potential of co-constructing sXAI. Traditional XAI methods, as discussed, often treat the explanation process as a one-way transfer of knowledge, assuming fixed explananda and equal treatment of all explainees. In contrast, co-constructing sXAI envisions a dynamic, interactive process in which the explanandum and the explanans evolve during the interaction, thereby accommodating the explainee's growing understanding and specific needs. This approach requires a mutual exchange in which the roles of explainer and explainee can interchange, fostering a more personalized and adaptive explanation framework. By focusing on the alignment of beliefs between the explainer and explainee, co-constructing sXAI aims to create more effective and context-sensitive explanations, advancing the field toward more human-centered AI systems.

3.7 Rapid Access to the Content of This Chapter

Explanations are not straightforward, static processes. They evolve with the understanding of the explainee and the context of the interaction. The dynamic nature of an explanation means that roles can interchange between the explainer and explainee, with an explainee possibly contributing knowledge that alters the original explanandum (subject of the explanation). This shift contrasts with traditional XAI (explainable AI) in which explanations are typically one way.

This chapter focuses on the core components of explanations in social explainable AI (sXAI) and their relevance to human–AI interaction. The essential elements of an explanation are identified, and how these elements interact is explored, especially in dynamic, evolving explanations. The key concepts are the following:

- Explanandum: the object or entity being explained, such as an AI decision, event, or action
- Explanans: the method or content of the explanation, often co-constructed through interaction between explainer and explainee
- Explainer: the human or nonhuman agent in the role of producing the explanation
- Explainee: the human or nonhuman agent in the role of an addressee of the explanation, often the human in XAI scenarios

These components form the foundation for understanding how explanations can be structured in AI systems.

Role of Explainer and Explainee

The roles of explainer and explainee are dialogical. The explainee strives to understand an explanandum, and the explainer supports them in overcoming a knowledge

gap. Therefore, the explainer typically knows more about the explained domain than the explainee. Explainer and explainee co-construct the explanation together. As the explainee's understanding grows, the explainer adapts the explanation. In XAI, the AI classically serves as the explainer, with humans as explainees. Sometimes, however, these roles may reverse, such as when the explainee (human) provides feedback or additional information that alters the course of the explanation.

Explanandum as a "moving target"
The explanandum (the subject of explanation) is fluid, often changing during the explanation process. It adapts to the explainee's growing understanding or evolving needs. For example, in the context of sXAI, the explanandum might be the outcome of a collaborative task between a robot and a human or it might be an AI's decision that might—over the course of the explanation—change to the impact of an alternative decision on the explainee's life.

Explanans and Its Importance
The explanans (how an explanation can be verbally or nonverbally expressed and co-constructed by both partners) is tightly linked to the explanandum. For instance, the explainer may use a specific reference to existing knowledge to help the explainee understand the explanandum. The explanation strategy used by the explainer determines how the explanandum is broken down into smaller parts, influencing the effectiveness of the explanation. In sXAI, this can involve the AI system using natural language, visual cues, or multimodal feedback to explain its actions. The complexity of the explanans depends on factors such as context, environment, and interaction norms that affect how the explanation is structured.

Future Directions
This chapter suggests moving beyond current state-of-the-art XAI methods that assume fixed explananda and one-way knowledge transfer. Co-constructive sXAI envisions a more adaptive system in which explanations evolve based on the explainee's goals and understanding. This allows for more personalized, effective interactions. By focusing on how explanations are co-constructed and how explainer and explainee roles can adapt to each other, co-constructive social XAI offers a pathway to more human-centered AI systems.

Acknowledgments This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): TRR 318/1 2021 – 438445824. Rachid Alami's work has been partially supported by the EU-funded project euROBIN under grant agreement no. 101070596 and by the Artificial and Natural Intelligence Toulouse Institute (ANITI) funded by the France 2030 program under the grant agreement no. ANR-23-IACL-0002.

References

Alpsancar, S., Buhl, H. M., Matzner, T., & Scharlau, I. (2024). Explanation needs and ethical demands: Unpacking the instrumental value of XAI. *AI and Ethics* 1–19. https://doi.org/10.1007/s43681-024-00622-3

Arnold, T., Kasenberg, D., & Scheutz, M. (2021). Explaining in time: Meeting interactive standards of explanation for robotic systems. *ACM Transactions on Human–Robot Interaction, 10*(3), 25. https://doi.org/10.1145/3457183

Bokulich, A. (2011). How scientific models can explain. *Synthese, 180*, 33–45. https://doi.org/10.1007/s11229-009-9565-1

Buschmeier, H., Buhl, H.M., Kern, F., Grimminger, A., Beierling, H., Fisher, J., Groß, A., Horwath, I., Klowait, N., Lazarov, S., Lenke, M., Lohmer, V., Rohlfing, K.J., Scharlau, I., Singh, Terfloth, L., A., Vollmer, A.-L., Wang, Y., Wilmes, A., & Wrede, B. (2025). Forms of understanding of XAI-explanations. *Cognitive Systems Research, 94*, 101419. https://doi.org/10.1016/j.cogsys.2025.101419

Clodic, A., Pacherie, E., Alami, R., & Chatila, R. (2017). Key elements for human–robot joint action. In R. Hakli & J. Seibt (Eds.), *Sociality and normativity for robots* (pp. 159–177). Springer. https://doi.org/10.1007/978-3-319-53133-5_8

Garfinkel, A. (1981). *Forms of explanation. rethinking the questions in social theory*. Yale University Press.

Hellström, T., & Bensch, S. (2018). Understandable robots – What, Why, and How. *Paladyn, Journal of Behavioral Robotics, 9*(1), 110–123. https://doi.org/10.1515/pjbr-2018-0009

Holler, J., & Bavelas, J. (2017). Multi-modal communication of common ground: A review of social functions. In R. B. Church, M. W. Alibali & S. D. Kelly (Eds.), *Why gesture?* (pp. 213–240). Benjamins. https://doi.org/10.1075/gs.7.11hol

Human, S., & Watkins, R. (2023). Needs and artificial intelligence. *AI and Ethics, 3*(3), 811–826. https://doi.org/10.1007/s43681-022-00206-z

Johnson-Laird, P. N. (2005). Mental models and thought. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 185–208). Cambridge University Press.

Kern, F. (2007). Prosody as a resource in children's game explanations: Some aspects of turn construction and recipiency. *Journal of Pragmatics, 39*(1), 111–133. https://doi.org/10.1016/j.pragma.2005.01.017

Kern, F., & Selting, M. (2020). Conversation analysis and interactional linguistics. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Wiley. https://doi.org/10.1002/9781405198431.wbeal0203.pub2

Kim, J., & Im, I. (2023). Anthropomorphic response: Understanding interactions between humans and artificial intelligence agents. *Computers in Human Behavior, 139*, 107512. https://doi.org/10.1016/j.chb.2022.107512

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.

Lemaignan, S., Warnier, M., Sisbot, E. A., Clodic, A., & Alami, R. (2017). Artificial cognition for social human–robot interaction: An implementation. *Artificial Intelligence, 247*, 45–69. https://doi.org/10.1016/j.artint.2016.07.002

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

Pitsch, K., Vollmer, A. L., Fritsch, J., Wrede, B., Rohlfing, K., & Sagerer, G. (2009). On the loop of action modification and the recipient's gaze in adult–child interaction. In *Proceedings of the Gesture and Speech in Interaction International Conference*. https://www.honda-ri.de/pubs/pdf/1306.pdf

Robrecht, A., Buhl, H., & Kopp, S. (2024). Inferring partner models for adaptive explanation generation. In *Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue*.

Rogoff, B. (1998). Cognition as a collaborative process. In D. Kuhn & R. S. Siegler (Eds.), *Handbook of child psychology: Vol 2. Cognition* (pp. 679–744). Wiley.

Rohlfing, K. J., Cimiano, P., Scharlau, I., Matzner, T., Buhl, H., Buschmeier, H., Grimminger, A., Hammer, B., Häb-Umbach, R., Horwath, I., Hüllermeier, E., Kern, F., Kopp, S., Thommes, K., Ngonga Ngomo, A.-C., Schulte, C., Wachsmuth, H., Wagner, P., & Wrede, B (2021). Explanation as a social practice: Toward a conceptual framework for the social design of AI systems. *IEEE Transactions on Cognitive and Developmental Systems, 13*(3), 717–728. https://doi.org/10.1109/TCDS.2020.3044366

Sakai, T., & Nagai, T. (2022). Explainable autonomous robots: A survey and perspective. *Advanced Robotics, 36*(5–6), 219–238. https://doi.org/10.1080/01691864.2022.2029720

Schmid, U., & Wredem, B. (2022). What is missing in XAI so far? An inter- disciplinary perspective. *KI-Künstliche Intelligenz, 36*(3), 303–315. https://doi.org/10.1007/s13218-022-00786-2

Vollmer, A.-L., Mühlig, M., Steil, J. J., Pitsch, K., Fritsch, J., Rohlfing, K. J., & Wrede, B. (2014). Robots show us how to teach them: Feedback from robots shapes tutoring behavior during action learning. *PloS One, 9*(3), e91349. https://doi.org/10.1371/journal.pone.0091349

Woodward, J. (1979). Scientific explanation. *The British Journal for the Philosophy of Science, 30*(1), 41–67. https://doi.org/10.1093/bjps/30.1.41