Explanation as a social practice: Toward a conceptual framework for the social design of AI systems

Katharina J. Rohlfing, Philipp Cimiano, Ingrid Scharlau, Tobias Matzner, Heike M. Buhl, Hendrik Buschmeier, Elena Esposito, Angela Grimminger, Barbara Hammer, Reinhold Häb-Umbach, Ilona Horwath, Eyke Hüllermeier, Friederike Kern, Stefan Kopp, Kirsten Thommes, Axel-Cyrille Ngonga Ngomo, Carsten Schulte, Henning Wachsmuth, Petra Wagner, and Britta Wrede

***Abstract*—The recent surge of interest in explainability in artificial intelligence (XAI) is propelled by not only technological advancements in machine learning, but also by regulatory initiatives to foster transparency in algorithmic decision making. In this article, we revise the current concept of explainability and identify three limitations: passive explainee, narrow view on the social process, and undifferentiated assessment of understanding. In order to overcome these limitations, we present explanation as a social practice in which explainer and explainee co-construct understanding on the microlevel. We view the co-construction on a microlevel as embedded into a macrolevel, yielding expectations concerning, e.g., social roles or partner models: Typically, the role of the explainer is to provide an explanation and to adapt it to the current level of understanding of the explainee; the explainee, in turn, is expected to provide cues that guide the explainer. Building on explanations being a social practice, we present a conceptual framework that aims to guide future research in XAI. The framework relies on the key concepts of monitoring and scaffolding to capture the development of interaction. We relate our conceptual framework and our new perspective on explaining to transparency and autonomy as objectives considered for XAI.**

***Index Terms*—Explainability, process of explaining and understanding, explainable artificial systems**

I. Introduction

EXPLAINABILITY as a topic has recently experienced a surge of interest, even though it has been at the core of artificial intelligence since the start. It expresses the desire to make a system's behavior intelligible and thus controllable by humans (e.g., [1]). Two impulses seem to have been crucial for this recent interest: One comes from a technological perspective driven by the development of multilayered connectionist AI systems whose predictions (e.g., in medicine or jurisdiction) concern human lives; with their many nested layers and nonlinearities, machine-learned models have become opaque not only for citizens but also for experts [2]. This is especially threatening in the face of the mistakes and biases of deep learning systems ([3,4]). Opacity is "a serious issue in all those contexts where human beings are liable for their decision" [5, p. 5]. The concern to break open "black-box" algorithmic decisions has been addressed in regulations issued by the European Union (GDPR: General Data Protection Regulation)—the other impulse for explainability research. These regulations grant citizens a basic right for algorithmic decision making to be made transparent. The objective of making algorithms (or a part of them) accessible is at the core of eXplainable Artificial Intelligence (XAI), in which transparency, interpretability, and explainability are discussed as desired outcomes [6]. After recently reviewing the state of the art, Sokol and Flach [7, p. 235] concluded that "while a variety of interpretability and explainability methods is available, none of them is a panacea that can satisfy all diverse expectations and competing objectives that might be required by the parties involved."

Our article takes this conclusion as a starting point. Following [7], we argue that one important source from which XAI can tap diverse expectations is the interactive process of explaining. In this process, the receiver of an explanation does not just play a passive role of providing a set of properties according to which an explanation needs to be "personalized." Instead, in a truly interactive process, both partners – the explainer and the explainee – are regarded as social agents who not only have individual goals, intentions, and expectations but also construct these and agree on these jointly within the process. This construction allows the partners to engage actively, thereby intertwining the process of explaining with the process of understanding. However, accounting for this kind of dynamics requires the formulation of a conceptual framework.

In this article, we present a conceptual framework that allows us to study explainability as a social and interactive process. It addresses three main limitations that arise from recent research: the first limitation (Section IIA) arises from explanations typically being conceptualized as complete when they accurately describe the internals of a system. With its focus on the content of an explanation, this conceptualization takes little account of a receiver.

We argue that explanations should put the explainee and explaineer in the focus rather than only the properties of the explanandum. The second limitation is emerging in recent discussions in which scholars emphasize the need to personalize explanations (Section IIB). We argue that this need

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2020.3044366, IEEE Transactions on Cognitive and Developmental Systems

comprises more than an adaptation to the personal preferences or traits of an individual. For an explanation to be successful, the recipient's level of and progress in understanding also have to be taken into account. The final limitation, which we consider in Section IIC, concerns the knowledge gap (or: explanandum) that an explanation targets. Commonly, this is viewed as being identifiable prior to the interaction and as being fixed. In contrast, we argue that identifying/agreeing on the knowledge gap is itself an outcome of the interaction.

Our answers to these three main limitations guide us toward the framework (Section III) that emphasizes the interactive process and is founded on research on the following aspects of interaction and development: co-construction, monitoring, scaffolding, and social practice. This approach paves the way to reach the objectives of transparency and autonomy that are called for in research on XAI.

II. Motivation: Current limitations to the concepts of XAI

Our culture highly values explaining, both politically and individually. Applied to AI systems, the call for explainability responds to the current situation in which intelligent software continues to make often incomprehensible decisions that affect human lives [5]. In our society, there seems to be a consensus that such AI outputs have to be explained (or to be explainable) (see, e.g., GDPR, DARPA).

In this section, we review approaches that respond to this consensus and develop explainable AI. We inspect their underlying notion of explanation in order to reveal their limitations and shortcomings. Such an analysis of concepts is helpful: It reveals how the vocabulary used in the literature on XAI evokes specific ideas about how the relevant phenomenon can be formalized and modeled [8], and it helps to identify important next steps for the design of future AI systems.

A. A complete explanation is not enough

In the past, the objectives of research in the area of explainability have been connected to concepts of interpretability and completeness. In the following, we will first regard the two concepts before we then turn to explainability. Interpretability is defined as a description of "the internals of a system" [9] and often (but wrongly) equated with explainability (see, e.g., [10,5], for some corrections). Interpretable models offer procedures to simplify or inspect the output of complex systems. In Rosenfeld and Richardson's [11] terms, they focus on the question *what* should be explained. Completeness, in turn, captures the vision of a generic description that is understandable on its own, because it describes "the operation of a system in an accurate way" [9].

The limitations of both concepts become visible when considering the basic terms relevant to explainability: The *explainer* is the person (or a system) who explains and the *explainee* is a person (e.g., an adult, a child, a learner, or a learning system) addressed by this explanation (see Fig. 1). The object of explanation is the *explanandum* (e.g., [12]). The term *explanans* refers to the way in which the explainer conceptualizes the object of explanation (see Fig. 1).

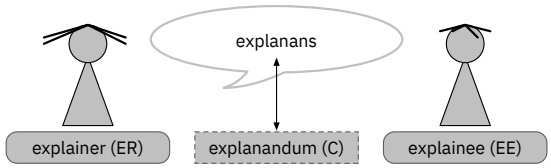

Figure 1. The main elements of an explanation.

The notions of interpretability and completeness build on a range of implicit assumptions such as: (a) The explanandum as the object of explanation (see also Fig. 1) exists in the world independently of the interaction between the explainer and the explainee. It is further assumed that (b) the mere availability of information will suffice to (c) warrant an understanding that (d) enables the explainee to act further. Clearly, the explainee is "conceptualized as an independent actor, who makes autonomous decisions on the basis of information made available to them through transparency" [13, p. 9]. Thus, the explanation serves the purpose of providing information that is independent of the explainee and her or his desires, goals, or social roles. This makes the explainee exchangeable. However, models based on interpretability and completeness need to be contrasted with explainable models concerned with the explainee's understanding [5,9,11]. In contrast to interpretability and completeness, explainable models are concerned with how to make a description understandable to the explainee. In other words, in addition to the *what* question (being at the fore of research on interpretability and concreteness), they also address the *why*, *who*, *when*, and *how* questions that also need to be answered, because they "heavily affect" the generation of an explanation [11, p. 696]. In this vein, "[e]xplainable models are interpretable by default, but the reverse is not always true" [9].

The advantage of explainable models is, thus, that they take a broader view of explanation and regard the explainee as a part of it. However, whereas current approaches – as we will argue in the next section – take the diversity of explainees in terms of their expectations into account [10], they barely consider the fact that an explainee is not a passive receiver. As a social agent, she or he can steer the explanation process significantly, as we exemplify in Table 1. At this point, while we second approaches calling for more attention to social aspects [10], we seek to go beyond an acknowledgment that persons employ social expectations when generating or evaluating explanations. We are interested in when such expectations are manifested to modulate the interaction, and how they change during a process of explaining. Such changing expectations, however, are currently not taken into account in XAI.

B. Social interaction is not just about personalization

Following the call to extend XAI to social and interactive approaches, novel developments in XAI acknowledge the diversity in explainees in terms of their expectations, interests, and needs as a way to personalize explanations [14,15,11,16,7]. These approaches characterize the explainee in terms of a number of variables that specify the person's characteristics

(e.g., social role, personality, motivation or expertise, and circumstances such as cognitive load and processing time, etc.), and base the explanation process on this assessment (e.g., [17], [18]). They link up with a line of research that has widely recognized the partner's understanding as being central to a successful interaction. To ensure the partner's understanding, models of the partner have been proposed for the interaction in general [19,20]. Computational approaches to such partner models (often called *user models*) have been defined as "knowledge sources [containing] explicit assumptions on all aspects of the user that may be relevant to the dialog behavior of the system" [21, p. 6]. These factors can be processed before an interaction commences [22].

To implement such a personal context of a user, a system has to be 'aware' of what kind of partner and situation it is facing and choose an action that seems appropriate to its situation analysis—that is, user models represent dialog- and discourse-related aspects of the interaction (e.g., the common ground) as well as task-related (and task-specific) aspects (e.g., the user's knowledge level) [18]. Depending on the user's configuration, an explanation can be evaluated in terms of its adequacy [15].

User models are commonly rooted in popular theoretical approaches characterizing persons in terms of their mental states (in forms of goals, intentions, and desires). A recent review of explainable agents reveals that the majority of current XAI studies concerned with explainable and intelligent agents frame their objectives with concepts related to Theory of Mind [6]. Theory of Mind refers to the ability to attribute intentionality to an agent [23], often described as "mindreading" [6, p. 1079]. The main points of this theoretical perspective concern not only the question whether persons attribute intentionality to other beings, but also how they exchange information about mental states. In many studies, the exchange of mental states falls short, because it focuses on the moment in which a mental state can be read in the partner. The full function of exchange, however, implies that in both partners, the states can change, become aligned, or diverge from each other as the interaction proceeds (see Table 1 for an example)—properties that are barely in focus.

Counter to this state of the art, developers of AI systems focus on social interaction being about 'reading' the intentions of the partner in general. Applied to the process of explaining, this implies that to know what the addressee wants or needs is at the core of a successful explanation [9]. However, the fact that the adaptation process should go beyond the first impression receives little attention, despite being acknowledged in literature on interaction (e.g., [24]). In this respect, we argue that the explainee should be characterized by a complex and dynamic variable that captures not only personal traits but also the explainee's continuously changing level of understanding. In other words, the extension of XAI research to social and interactive approaches (as suggested by [10] and [7]) should not limit itself to characterizing an explainee as an individuum with preferences, personal characteristics, intentions, etc. Instead, it is necessary to account for the explainee seeking to understand the explanandum. Putting this role at the center requires methods to dynamically model the negotiation of information, level of understanding, and changing knowledge.

Currently, hardly any framework exists that could be used to account for such dynamics in which both partners, as social agents, are actively engaging in and shaping the process of explaining. The premise to involve both partners stands in clear opposition to existing frameworks for explanations in AI (e.g., [5]). In other words, our approach differs from approaches targeting just one comprehensive and personalized explanation that is configured along some parameters such as the why, who, what, when, and how [11]. In our approach, the target is to allow an explanation to be configured or modified *within* an explaining process in which the explainee co-constructs the explanandum in order to arrive at an understanding.

The involvement of both partners is crucial to our approach and is based on research regarding the dialog as a unit of analysis. Accordingly, interaction does not just serve the coordination of individual actions but actually gives rise to *joint actions* that are of a different nature than actions performed by an individual for a noncollaborative purpose [20,19,24,25,26]. Those collaborative actions result in different ways of processing the physical world—an effect that is already visible in infants [27]. Collaborative actions request partners to align [28,29] and to design their behaviors for each other [30]. Applied to the explaining processes, the dynamics involve processes of monitoring the understanding that crucially steer the interaction. Below, we will show that such dynamics are important to consider, especially when it comes to everyday explanations.

C. *Scientific explanations are not everyday explanations*

Up to now, explainable AI research has focused mainly on scientific explanations—that is, on accurate and complete explanations of a model of some phenomenon. In contrast, everyday explanations can be understood as partial descriptions of the causes of some phenomenon that enhance understanding in the user in terms of why something happened (cf. [31]). Miller [10, p. 3] perceives important differences between scientific and everyday explanations. In contrast to scientific explanations that aim at exactness and completeness to fill a knowledge gap, the process of explaining in everyday language use is less driven by specific information needs. Instead, knowledge gaps can emerge in an explainee during the interaction. In addition, they are likely to change as the explainee might be able to clarify something while realizing novel knowledge gaps. Focusing on everyday explanations thus implies modeling the dynamic process of understanding, and the knowledge gaps can be addressed by simultaneously being a byproduct of the interaction [32, p. 229]. In this way, the explanandum becomes a 'moving target' rather than being defined a priori, as is the case for scientific explanations.

Adding to the dynamics, a crucial characteristic of everyday explanations is the variety of types. In fact, Kotthoff [33, p. 121] observed that in everyday life, types of explanations range from instruction giving to storytelling. The diversity in types occurs concomitantly with *various forms of understanding* that are

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2020.3044366, IEEE Transactions on Cognitive and Developmental Systems

typical for everyday explanations [32,34]. In this vein, the everyday nature of explanation might account for why, under some circumstances, a shallow explanation can be more successful than an elaborated one; and why, in other cases, even elaborated explanations fail. In contrast to scientific explanations that target a concrete phenomenon, everyday explanations require more flexibility in what content to capture and what form of understanding is necessitated (see Table 1). Research in AI has hardly focused on such flexibility. The development of ad hoc views and relationships that can be generated in the moment as part of an explanation could therefore be a future issue for XAI systems.

TABLE 1
INTERTWINING EXPLAINING WITH UNDERSTANDING

Aspects	Brief description
Example: Imagine a person searching for a new job. Optimizing the qualification structure of a cohort by considering clients with similar characteristics, a deep-learning algorithm provides a very unusual combination of further qualification as being the best available strategy.	
A Constituting explanandum	Depending on the dialog, the reasons for the suggestion, the way how to deal with it (accept or reject it), or any further signal might be crucial in constituting the explanandum; the explanandum is thus a '**moving target**.'
C Form of understanding	The kind of understanding needed **is difficult to foresee**: The person searching for a job could desire a justification for the suggestion to critically call it into question or she or he could also require more guidance about further steps.
D Dynamics of understanding	Understanding **can change** during the process of explanation: After being introduced to the professional area, a person searching for a job might understand more about, e.g., health services. In addition, her or his need to understand can change due to increasing knowledge.
E Roles in explaining/ understanding	The output of the system goes hand in hand with **social roles in a dialog**: On a microlevel, there is a person who asks and a system that delivers answers. On a macrolevel, the system providing the suggestions is more knowledgeable than the person.
F Monitoring/ Scaffolding understanding	A system can monitor and scaffold the person's understanding by developing a task-specific **partner model** that is derived from previous experiences: At any time in the dialog, this model needs to be adaptable.

To sum up the above-mentioned limitations, current research in XAI lacks a conceptual framework that would account for both explaining as a bidirectional social process and the dynamics of understanding in everyday explanations. A conceptual basis that allows one to assess and describe the process of explaining could enhance the design of AI systems tremendously [7] by orienting them toward the production of socially relevant explanations that cover not only specific points of interest in the addressee, but also the general dynamics of the process taking place on different levels. A conceptual basis is necessary, because in a recent review, Anjomoshoae et al. [6, p. 1082] revealed that 39% of research concerned with explainable and intelligent agents "did not rely on any theoretical background related to generating explanations"—thereby strongly suggesting that current theories might be lagging behind what designers of AI systems already recognize as being more appropriate.

A solid conceptual basis requires one to account for the asymmetry in the interaction that is to be found prior to an explanation. Furthermore, a conceptual basis also requires empirical evidence to support it. Below, drawing from linguistics, psychology, and developmental studies, we propose a conceptual framework that attempts to capture the social process of explaining.

III. Conceptual framework

Our goal is to propose a conceptual framework that allows us to study explanation as a social process and overcome the three limitations mentioned above. This framework, we claim, is useful for studying adaptation beyond personalization and within an explanation as a social process in which the explainee and explainer interact in different social roles. Finally, this framework allows us to move beyond scientific explanations and depart from the assumption that the goal of explaining is to deliver a complete and accurate explanation of an a priori defined and fixed explanandum. Instead, we focus on the incremental and interactive construction of the explanandum as a result of the interaction.

TABLE 2
DYNAMIC ASPECTS OF THE EXPLAINING PROCESS

Concept	Brief definition	Level
Co-construction	Bidirectional interaction involving both partners in constructing the task and its relevant (recipient-oriented) aspects	Micro
Monitoring	A key mechanism of a social interaction enabling partners to align and jointly act	Micro
Scaffolding	A form of assistance from a more knowledgeable partner who is adjusting the task to the learner's abilities	Micro
Social Practice	Established (but flexible) interaction pattern consisting of joint actions toward a goal that are performed in a sequence	Macro

A. Explaining in basic terms

In this section, we aim to propose some extensions to the basic terminology of explaining that will address the novel dynamic aspects (see Table 2).

We start by pointing to specific social roles in a dialog that are fulfilled by at least two persons interacting with each other for the purpose of resolving a factual or anticipated epistemic asymmetry [35]: the explainer and the explainee (see Fig. 1). The social roles become manifested as dialogical roles for which some interactive behaviors are characteristic. For example, in a tutoring dialog, tutors were found to ask questions which tutees are expected to answer [36]. Across disciplines, explanans and explanandum are at the center of explainability research in which they are designed to reveal the causes and relationships underlying a phenomenon. Whereas in current linguistic and psychological research, the explanans is limited to verbal means and little is known about how the explanandum can be expressed nonverbally, most current explainability research in computer science focuses on visual approaches to explanation by using, for example, Shapley values [37], individual conditional expectation [38], local surrogates [39], or other similar techniques (see [40], for an overview). An increasing number of works in computer science, however, do target verbal explanations by using verbalization techniques for

formal languages such as OWL [41] or even target the use of conversational agents based on class-contrastive counterfactual statements [42]. These approaches value the fact that visible aspects are often insufficient to stand on their own and need to be framed verbally in order to foster specific forms of understanding. Multimodal explanations are currently limited to visualization combined with textual explanations [43]. These approaches depend highly on the user's expertise and established vocabulary that form a basis for revealing relevant underlying causes and relationships.

As already mentioned, the terms have evolved from research on scientific explanations and refer to an explainee experiencing a specific knowledge gap [44]. However, in natural interactions, the explanans and explanandum might be underspecified or not defined at the beginning of an interaction. In this vein, the object of explanation might not exist before the interactional exchange (see also Tables 1 and 2). The aim of our conceptual framework is, thus, to capture the dynamics of the explaining process along aspects identified in Table 2. Accordingly, explaining is regarded as a social practice that is co-constructed through constant monitoring and scaffolding by both explainer and explainee. In the following, we will elaborate further on each of the aspects and show how they are connected with each other on different levels.

B. *Co-constructing understanding*

Table 1 shows that the reaction of the explainee is difficult to foresee; and that for AI systems, everyday explaining requires a flexibility to account for different forms of understanding. To account for the dynamics of everyday explanations, we need to change our view on the process of explaining as such and see that it does not – or does not only – comprise a unidirectional transfer of information from one person to another. Instead, explaining is a bidirectional and iterative process [45] in which humans implicitly or explicitly negotiate and construct the explanandum, the explanans, and their form of understanding. In other words, when it comes to everyday interactions, explaining does not just 'tease out' a specific form of understanding that the explainee already possesses. Importantly, understanding is an interactive and constructive process [19,20,46,47].

In proposing a novel conceptual framework, we claim that the key aspect that captures the dynamics of everyday explanations is *co-construction*. This refers to the process by which both interaction partners, the explainer and explainee, construct an explanation in close relation to not only the emergent understanding but also their broader knowledge, values, and assumptions.

A byproduct that results from a co-construction is **a context** that renders an explanation relevant for both participants. The construction of such a context can be achieved by a partner model (described above). Such a context can be described as a "collection" [48, p. 22] of material, social, or physical facts that need to be taken into account when persons interact—this is a view that is currently common not only in computer science but also in psychology and cognitive science in general.

In contrast to a loose collection of facts, we view an explanation as operating on (at least) two levels: micro and macro. Whereas the microlevel unfolds during an interaction, the macrolevel releases the context from the material or social environment(s) and makes it more dependent on the shared knowledge on which the partners will agree [48]. The advantage of such "emergent context parameters" [48, p. 22] is that they are flexible, highly relevant, and can thus be constructed on demand. As of yet, we are not aware of any computational approaches that use such a notion of context. Hence, in the following, we describe the two levels relevant for the co-construction of understanding.

IV. Microlevel of Co-construction: Monitoring and Scaffolding

In this section, we turn to the question of how to account for the dynamics of everyday explanation by arguing that what is required is a focus on the *process* of explaining and understanding.

Concerning the microlevel, studies on natural interaction involve the concept of "common ground"—that is, mutual understanding that is established between interaction partners [50, p. 127]. The phenomenon of a context is well-researched, based on pragmatic approaches (see [49] for an overview), and implemented in computational models of dialog [51,52,53,54]. Yet, the mechanisms of common ground (e.g., how it is managed, how it is represented, and how it is influenced by different joint goals) are still subject to discussion in psychology and (psycho)linguistics (e.g., [55]). Relevant to XAI, the open questions center around the emergence of the common ground that is needed to arrive at a successful understanding. Below, we use two concepts to outline how, on a microlevel of interaction, co-construction can be achieved.

A. *Explaining is monitoring*

We now return to the various forms of understanding as a characteristic of everyday explanations. What forms of understanding can be co-constructed in explanatory settings with AI systems is currently not well known, and this research gap should certainly be filled when further developing XAI. Research on technological explanations posits two different forms that need to be considered [56]): a mechanistic information on the architecture and an interpretative information on the artifact's function/relevance. Mechanistic artifact explanations focus on what has been labeled scientific explanations—namely, on information about mechanistic aspects that can be objectively correct or not such as the mechanical and physical workings of a motor car. In order to explain how to drive a motor car, however, only those mechanisms need to be explained that are related to the intended function of this artifact. The technical function needs to be explained with respect to a use plan [56]—that is, the goal and purpose that the artifact was designed for and that is ultimately bound to its social use context. Consequently, it is entangled with social norms, individual or intersubjective goals, and so forth. In other words, this kind of necessary information is not

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2020.3044366, IEEE Transactions on Cognitive and Developmental Systems

an objective part of the artifact, but a social ascription representing its meaning (its purpose and relevance). Both the mechanistic information on the architecture and the interpretative information on the artifact's function/relevance need to be accounted for in XAI, especially when they are supposed to support the agency and autonomy of the individuals [57]. This, however, is an interdisciplinary challenge [58]. Facing this challenge, it is interesting to note that the dynamic process of co-construction also impacts on the explanandum that can change even within the process of explaining. For example, starting with a mechanistic view, an explainer might end up providing insights into the technical function. It is thus important to study cognitive operations for changing the perspective on the entity or phenomenon.

Facing the various forms of understanding, a central question is how can an explainer support the process of understanding. Surprisingly, there is currently little research addressing this topic in the development of XAI. In general, taking an interactive view, recent dialog theories suggest that as the interaction unfolds, partners take notice of each other and align to each other on various behavioral levels [59] when pursuing a joint goal [29]. In addition to theories focusing on the phenomenon of alignment [59,29], other theories emphasize that the goal of coordination is not to converge on internal representations but rather to accomplish an activity together [24]. In this accomplishment, *monitoring* plays a key role: It is the core mechanism by which partners perceive each other's behaviors in order to jointly determine the course of each utterance [20,30]. More specifically, speakers have been found to monitor for (visible) evidence of understanding. For example, Clark and Krych [20] demonstrated that dyads (i.e., pairs of interactants) who could not monitor each other at all made eight times as many errors as dyads that could take advantage of monitoring each other. Studies that have been performed on other than explanatory tasks have shown that the understanding displayed by the interlocutor [19,60,61,62,63] and the modalities via which it is expressed [64] are informative to the speaker when, for example, reformulating an utterance [51], adjusting the modalities [64], or addressing the satisfaction and motivation of the interaction partner [65].

Findings such as this led scholars to claim that a function of a conversation cannot be defined on the level of the individual [24]. In other words, functional organization makes sense only within the dyad—the partners complement rather than copy each other [24]. By monitoring, they work together toward understanding [19,46,64].

What holds for studies on interaction in general should be verified for an explanatory dialog in particular. What is particular to explanatory dialog is that an explainer not only pursues the goal of conveying knowledge to the explainee (who agrees on gaining this knowledge) but also has to monitor her or his progress. Because of the different social roles linked to a knowledge asymmetry, an explanatory dialog thus calls for a different organization of alignment in partners. We propose that alignment for the explanation-specific dialog can be captured by two operations:

- predicting the other's behavior, and accordingly
- conceptualizing the explanans.

Below, we specify how these two operations drive the explanans and the explanation process (see Fig. 1).

We have argued that the assumptions about what and how to explain might be a result of a former explicit negotiation of the explanandum, an implicit negotiation (also unexplored in research and linked to the history of interaction), but also the way in which the explainer conceives her or his social role within an explanation [62]. Here, the social practice of explaining (see Section V) will certainly provide a useful 'template' for how to explain from a macrolevel perspective. Socialized in routines, social practices impose, for instance, obligations on the participants [53] that are defined in terms of what is permissible (ibid). They create expectations that, in turn, can be imparted to a partner model. Within this frame, an explainer can undertake further steps in her or his social role to guide the explainee toward the desired outcome. Hence, monitoring as a process clearly intertwines the micro- with the macrolevel.

Whereas our view on explaining as a social practice (see below) leads us to propose an interactional structure, other authors find a mentalistic structure to stimulate the explanatory dialog [66]. Be it interactional or cognitive, only within such a frame and by monitoring the explainee's progress, can the explainer *predict* the partner's behavior, estimate its appropriateness, and monitor the progress of understanding—a solid basis for helpful feedback. Conceptualizing the explanans is a consequence derived from this solid basis. It is the ability to *formulate* the subsequent action of explaining on the basis of whether or not the actions (involving forms of understanding) of the explainee correspond to the predicted actions (cf. [59]). In other words, when conceptualizing explaining, the explainer links her or his actions to the ongoing situation and interaction by modifying the assumption about how and what to explain. In this respect, we propose that multimodal signals of understanding, partial understanding, nonunderstanding, or misunderstanding that can be monitored provide a scaffold to the explainer in the sense of an impulse to adjust or terminate the explanation. It is through this that the explainee actively shapes the explanans (cf. [61] for tutoring).

Because an explanation is a practice that serves the goal of solving an epistemic gap, there is, at some point, pressure on the explainee to demonstrate an understanding [67]. This pressure might be less when interacting with an artificial system. As Howley and colleagues [68] have shown, students ask more questions when learning with a robot when compared to learning with a human teacher.

B. Explaining is scaffolding

Scaffolding behavior has been observed in task-oriented developmental studies on adult–child interactions. In these asymmetric interactions, adults contingently provide support based on the child's performance as well as on her or his cognitive and linguistic abilities by, for example, increasing their assistance to the less competent child [66]. As a key aspect

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2020.3044366, IEEE Transactions on Cognitive and Developmental Systems

of social interaction, contingency captures the timing of social interaction (e.g., [69]), whereas scaffolding refers to the way a contingent support is formulated in an asymmetric interaction. The function of scaffolding is twofold: On the one hand, it enables "a child or a novice to solve a problem, carry out a task or achieve a goal which would be beyond his unassisted efforts" [66, p. 90]. On the other hand, the assistance concentrates upon "those elements that are within the learner's range of competence" (ibid). Against this background, a scaffolding behavior thus requires (a) a mental decomposition of a targeted action and (b) enriching the obvious and visible to the learner with aspects that the learner either cannot discover by her- or himself easily or that enrich the perceivable events in order to better interpret them. In this respect, Wood et al. [66, p. 97] already proposed that in a scaffolding process, a tutor needs to have "a theory of the task or problem and how it may be completed." The goal in scaffolding is to allow a less competent partner to participate in an interaction and to contribute to a task. Participation can first be achieved with the contingent support of a more competent partner, but at the end of the learning process, it will be achieved independently and in a self-regulated way [70] Thus, an important feature of scaffolds is that they are temporary. The notion of scaffolding is special in that it does not simply emphasize support or assistance by a more competent partner who reduces the complexity of the learning content. It also emphasizes that learning is co-constructed by both partners: the learner, who signals her or his individual level of readiness [66], and the more competent partner, who adapts accordingly and provides support just above the level of learner's abilities [71].

For our conceptual framework, we transfer the concept from the area of learning to the area of explaining and extend its definition to everyday explanations that require understanding to be co-constructed: Consequently, both partners can scaffold each other—that is, one partner can provide the other partner with the additional information needed to arrive first at the explanandum and then at the goal of explanation: namely, the understanding in its different forms.

In our example (see Table 1), a system can scaffold the person's understanding by developing a task-specific partner model that is derived from previous experiences with, for example, an explanandum, a particular explainee, or a specific dialogical role. Such a partner model generates hypotheses about how to explain and how to produce the explanans to converge on the learner's understanding (cf. [66]). It becomes manifest in a multimodal modification of interactive behavior (e.g., [30,64,65,72,73]). The helpful function of such models has already been recognized by Cawsey [22], who suggested that a particular design of an explanatory interaction needs to take place in accordance with the user's knowledge, the current object in focus, and the role of the participants—all factors that can be determined beforehand (cf. [74]). "Depending on the user's assumed current knowledge, different explanation strategies will be selected, prerequisite information either included or left out" [22, p. 6].

What is informative for our approach is that a scaffold in the form of an explainee's model prior to an interaction then needs to be verified, refined, or modified in that interaction. The impulse for the adaptation of the scaffolding behavior comes from the goals set by the explainer, but, most importantly, from analyzing the reaction of the explainee. For example, by observing how parents manage a moment in time when the attention of children was vanishing from the demonstrated action, Pitsch and colleagues [63] found that parents produced larger motions to regain their children's focus. Thus, there seem to be some multimodal (verbal and nonverbal) parameters that can be modified when the interaction affords it [30,64]—some of them connected to the perception and some to the task structure [75]. Again, we lack empirical findings on what kind of scaffolding mechanisms are active in explanatory dialogs and what kind of parameters steers them. Yet, such an adaptation process, we argue, is necessary to customize an explanation to the explainee. Without it, an explainer can neither generate relevant feedback nor devise dialog in which her or his feedback will be more appropriate for *this* explainee in *this* task at *this* point in the process of understanding [66, p. 97,11].

Whereas above, we have focused on the explainer, according to the idea of co-construction, the explainee also participates in the scaffolding process. Below, we indicate how the two partners are coupled.

To summarize the concepts that we consider necessary on the microlevel, we propose that an explaining process is co-constructed between the interaction partners and consists of scaffolding and monitoring. Both processes on the microlevel of an asymmetric interaction characterize the dynamics needed and contribute the first part of why persons consider an explanation to be relevant and successful. At this stage, we propose the following formalization of our framework (Fig. 2):

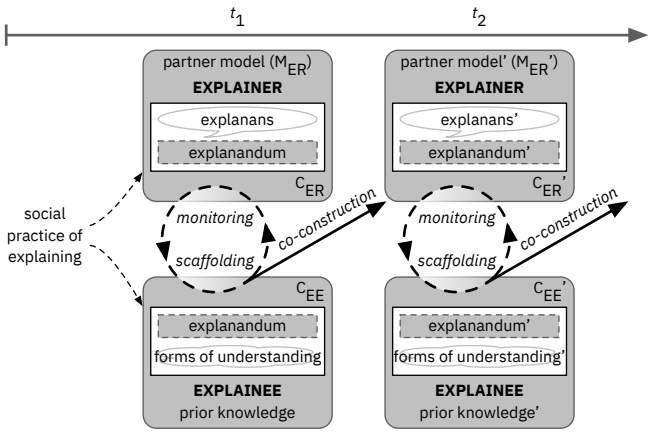

Figure 2. Our approach to **co-constructing the explaining process**: There are two timepoints (t1, t2) that stand for an unfolding interaction across which partners adapt their behavior by monitoring and scaffolding each other. There is also prior knowledge from earlier interactions brought into this interaction. Both partners, the explainer and explainee, enact a social practice of explaining, the structure of which gives rise to specific social/dialogical roles, behaviors, and expectations.

We assume that the explainee (EE) has a certain understanding or conceptualization of the explanandum $C_{EE}[t]$. Note that C_{EE} is time-indexed, because the understanding of the explanandum

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2020.3044366, IEEE Transactions on Cognitive and Developmental Systems

will evolve or change over the course of the interaction. The explainer (ER) also has a certain understanding or conceptualization of the explanandum: C_{ER}. ER needs to monitor EE's level of understanding by comparing C_{EE} to the own conceptualization of the explanandum, C_{ER}. However, as the explainer cannot access $C_{EE}[t]$ directly, she or he interprets signals from EE to infer a model of $C_{EE}[t]$. We denote this inferred model of the level of understanding by EE with $M_{ER}(C_{EE}[t])$. We further assume that ER has a model of what she or he intends to explain to EE—that is, a model of what EE needs to understand. We denote this model of what the explainee should understand as $G_{ER}[t]$. Here, G stands for goal and is time-indexed as well, because the goal can change during the interaction. At each point in the interaction, the explainer is able to monitor how close $M_{ER}(C_{EE}[t])$ is to $G_{ER}[t]$ and take actions accordingly by scaffolding the explainee's understanding.

Given this framework, further research questions are:

- How can the emerging conceptualization about the explanandum be modeled?
- How does the explainer infer the model of the emerging conceptualization? By which signals does she or he know whether EE understood something?
- How does ER compute the difference between what she or he expects EE to understand, $G_{ER}[t]$ and how she or he estimates the understanding of EE in $M_{ER}(C_{EE}[t])$?
- How does ER react to the computed asymmetry or difference between $G_{ER}[t]$ and $M_{ER}(C_{EE}[t])$, i.e., $\Delta(G_{ER}[t], M_{ER}(C_{EE}[t]))$?
- How does ER modify the goals $G_{ER}[t]$ over time depending on the signaled level of understanding in EE?
- Which operations or mechanisms does ER apply to modify the explanation and to scaffold the EE's understanding as measured by $\Delta(G_{ER}[t], M_{ER}(C_{EE}[t]))$?

The limitation to our formalization is that is does not yet consider the macrolevel. The microlevel, however, is strongly informed by the macrolevel that we introduce next.

V. Macrolevel of co-construction: social practice

Whereas on the microlevel of interaction, the notion of context is recognized in the development of AI systems to some degree, context on the macrolevel is barely considered. Its formalization is clearly a topic for further research.

A. Explaining as a social practice

Miller [10, p. 51] suggests that explanations can take place only as a part of a conversation adapted to the explainer's and explainee's beliefs and oriented toward conversational routines. However, there are two additional aspects that, in our view, require consideration and further research.

The first addition concerns Miller's view [10, p. 6] on the process of explaining as "knowledge transfer." Although the transfer of knowledge is a key element within an explanation (see "Job 3" in Fig. 3), this element belongs to a larger structure. According to recent research in linguistics, explaining consists of a sequence of actions that is constituted by a specific goal (e.g., to fill the epistemic gap) and by the interactively accomplished conversational subtasks or "jobs" [76], [77, p. 84]. The way to perform such sequences provides a protocol (e.g., [78]) that serves as an orientation in the sense of what to do next and an interpretation of where an action could be going. Various protocols exist, according to which an explanation differs from, for example, an interpretation or an evaluation.

For the protocol of an explanatory dialog, Morek ([35]; see also Fig. 3) identified the following jobs: (1) establishing topical relevance; (2) constituting an explanandum; (3) explicating procedural, conceptual, or causal relations; (4) closing; and (5) transitioning to another talk.

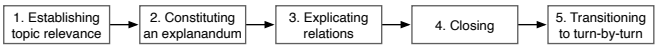

Figure 3. The jobs constituting the activity of explaining [35].

We view the protocol consisting of Jobs 1–5 as a context that unfolds for an explanation on a macrolevel. This structure modulates the interaction, because these jobs require the partners to coordinate according to a protocol (e.g., [78]). This coordination is achieved by using verbal and nonverbal resources (e.g., [79]). The protocol emerges because of established routines that – within the history of joint interaction [78] – have proved to be successful for achieving a goal. Research in linguistics characterizes such routines as *practices* (e.g., [79,80]). The notion of practices brings us to the second aspect that requires consideration from a macrolevel. Whereas Miller [10, p. 2] has pointed vaguely toward "certain biases" and "social expectations" that people make use of to generate or evaluate explanations, we propose that the biases and expectations emerge because explanations are practices. The concept of a social practice allows a proper contextualization of the interaction process [81] and yields macrolevel structures in the form of "social presuppositions" [82, p. 89] that go beyond the visibly performed actions: On a macrolevel, explanations are part of larger social settings (e.g., education, workplace, or job training) modulated by physical and social facts such as time to process, economic resources, subject positions, dialogical roles, capabilities to act, reputation, and so forth.

As already indicated, the macrolevel modulates the interaction on a microlevel. More specifically, it brings about new criteria for an explanans, because it has to be adequate with respect to, for example, the social status of the explainee. Thus, it is important to discern that any form of a protocol or communicative practice [83] does not begin with individuals but exists "prior to these individuals who are called upon to give it life" [84, p. 865]. Whereas these structures on a macrolevel are inevitably present when actions are performed in a collaborative way, on a microlevel, the ongoing interaction has the power to ratify or change it, as the typical action sequence consisting of jobs (see Fig. 3) is not rigid (e.g., [85]).

Social practices are studied on different levels of analysis in disciplines such as linguistics and sociology. Even though Matzner and colleagues [86] as well as Neyland [87] have established the relevance of such physical and social facts for

several applications of AI technology already, approaches toward XAI have not made tight connections to the respective social science research, but have been drawn mostly from single findings or coarse analogies (if not simply intuition, see [10]). With our additions to Miller [10], we are proposing concrete social structures generating expectations that, so far, have received little attention in XAI.

What is important in the example presented in Table 1 is whether the person looking for a new job receives a suggestion within an interaction (microlevel) that suits the person's dialogical role (macrolevel). In this sense, a social practice of explaining is not just a background or a set of conditions for the interaction. Instead, it determines the (social / dialogical) roles of the subjects involved (e.g., [80]) and how that explanation will be interpreted, although this interpretation can become modified during the course of interaction—this modification speaks to the fact that persons are actively involved in constructing (and reconstructing) the social practice [81,85].

Because the social practice within which an explanation occurs can change the social relations and power structures, any act of explaining also involves normative aspects.

For the design of explainable AI, the concept of social practice requires the implementation of representations from at least two levels: Based on social presuppositions on a macrolevel, a protocol can guide the explanatory dialog and specify how to act in accordance with whether, for example, an action or a concept has to be understood. In addition to the macrolevel, computational models also need to account for the construction/development of this structure depending on the interaction's course on a microlevel, its goals, and its means that go beyond verbal behavior to cover patterns of speech and gestural behavior or affect. This is the reason why representations need to be designed and implemented that will bind information from various sources and various levels (such as those prior to the process of explaining) not only during ongoing interaction but also across repeated explanations. Equipped with representations from at least two levels, a system can derive or adjust the context parameters from the process of interaction as we proposed above by formalizing it. Future research needs to investigate how these two levels are intertwined. For the design of artificial systems, one possible way to implement the two levels is in an emerging 'metasystem' that increasingly influences the interaction.

VI. The Macroconcepts of Transparency and Empowerment

We now return to the critical discussions propelled by the regulations issued by the European Union that are contributing strongly to the development of XAI. The GDPR claims that all persons have a basic right for algorithmic decisions to be made transparent. In this section, we discuss the relationship of the concept of transparency and empowerment to our framework.

Felzmann et al. [13] differentiate between different forms of transparency. Prospective transparency refers to the procedures by which users are informed about the data processing and the working of the system upfront. Retrospective transparency, in contrast, generates post hoc explanations and rationales. For an AI system to be retrospectively transparent, one should be able to inspect its "internals" to understand its decision. Instead of considering these different forms, as pointed out in Section II when referring to interpretability, AI systems seem to assume that a mere presentation of information can bring about understanding that enables an explainee to act further. These assumptions would be justified only if AI systems were to deal with persons who are literate in assessing the mechanisms and their consequences as well as the risks behind the data processing that the AI systems perform [13]. However, this is not the case, and we need to stress that these assumptions are not justified in general: McStay [88] already raised the point that there is hardly any basis according to which users could have a clear picture about mechanisms that process their private data and about the goals for which these data are processed. Along these lines, "to view individuals as rational economic agents who are able to go about deciding how to protect or divulge their personal information is highly misguided" [88, p. 599 f.].

According to our conceptualization of explanations, the goal of an explanation is to induce knowledge in the sense of practical understanding of an entity along with the capability of using it for a specific subsequent purpose (e.g., decision, learning, task accomplishment), thus, empowering explainees to act in an informed fashion. Clearly, conveying information, that is not to be equated with knowledge [89], is not enough as information can be processed differently by people, as they "differ in their ability to make use of information provided, and different types of information pose different barriers to understanding" [13, p. 5]. The aim to understand mechanisms – which has been in the focus of McStay [88] – can thus be considered as only one of many aims relevant to an explanation. In addition to the types of information that are relevant for one or the other stakeholder group, individuals also differ within one group.

Thus, understanding means clearly more than an access to information, because it connects with further actions. In this sense, 'completeness of information' is not the main goal of an explanation (see [33]), and more information is not always better than less information. Instead, in fostering knowledge and understanding, it is important to monitor the explainee's progress and to provide information accordingly. Relevant explanations are situationally tailored to answer the explainee's purpose and empower individuals to act within a certain social context and set of practices. Partial knowledge, thus, may allow sufficiently accurate but rapid behavior.

To sum up, current discussions taking the value of transparency into account appear to be limited to the claim to make the internals of a system accessible for whatever purpose. If, however, by entitling citizens in their knowledge about data processing and eventually to empower them (e.g., to adjust algorithmic or sociotechnical systems, at least to a certain degree, to their own needs and beliefs), transparency is meant to serve the increase of autonomy, then it should be

conceptualized from the perspective of the individuals demanding it [13]. We must be careful, however, to not add the empowerment of persons to a number of already existing burdens [86] in legal or ethical requirements.

Clearly, understanding resulting in empowerment is a value of explainability research. However, in everyday explanations, understanding seems to vary in its nature: It can take various forms ranging from deep to partial, from enabling a further action to comprehension of relationships and procedures. The form of adequate understanding has to be co-constructed in each case respectively to fit the explainee and her or his context of knowledge and actions. It seems that AI systems will need to dispose of adaptive representations that underlie any explanatory process in order to be able to co-construct the form of understanding.

VII. Conclusions

In addition to current research on explainability and seconding Sokol and Flach's [7] recently formulated call to customize explanations, we offer a conceptual framework for the design of explainable AI systems. With this framework, we postulate that explainable AI systems can generate highly relevant explanations that can guide the system "in a direction that helps to answer selected questions" [7, p. 239] when they act in an interactive and co-constructing manner. The interaction within which selected questions are answered can change both the course of an explanation and the explanation's content and should be seen against the background of an explainee's everyday understanding in its various forms. Whereas current research on explainability recognizes the need for social or interactive aspects [7,10], in this article, we focused on the process of explaining and were able to identify dynamic aspects of it on different levels.

On a macrolevel, explanations are a social practice—that is, a sequence of actions organized according to an interaction protocol addressing the asymmetry of communication and the goal of filling the epistemic gap (the explanandum). As a structure, this protocol gives rise to social presuppositions of an appropriate behavior pertaining to the expectations about the role in an interaction. As a practice, such protocols are already established but will be enacted each time on a microlevel along the sequence of actions that is guided by the goal that is being co-constructed and ratified continuously by the participants.

To achieve the particular goal of an explanation, we have proposed two mechanisms that influence the course of interaction on a microlevel: monitoring and scaffolding. Both are known from research on development and interaction, with scaffolding being at the core of social learning [66] and monitoring at the core of a successful interaction [20]. Scaffolding operates on a macrolevel, because there exist some initial ideas about the scaffolding process and how to fill the epistemic gap. Whereas these provide an orientation for the interaction, on a microlevel of an interaction, such ideas need to be formulated in an explanans and modified by closely monitoring the explainee's progress in understanding.

In sum, both mechanisms characterize the process of explaining as a joint endeavor toward a goal. Such an endeavor, we argue, can be implemented in explainable and interactive AI systems aiming at everyday understanding, because protocols exist detailing some of the stable parameters that constitute a social practice and are enacted frequently in everyday communications. However, despite their stability, social practice also brings the advantage of being flexible, because every interaction can be co-constructed between, and thus adjusted by, the partners. These properties – stability on the one hand and flexibility on the other – speak to a complex system that is formed between partners during the process of explaining. It is time to face this complexity and to approach interaction as a social dynamic system for a specific purpose. For this, we need a theory of co-constructing explanations and their underlying representations that will not only embrace the phenomenon, but also provide a good view over how, in such a system, understanding and trust develop. It will be interesting to see whether this new path of human–machine interaction results in citizens being entitled in their knowledge about data processing and eventually empowered to adjust algorithmic or sociotechnical systems to their needs.

References

[1] W. J. Clancey, "The epistemology of a rule-based expert system—Aframework for explanation," *Artificial Intelligence*, vol. 20, no. 3, 215–251, 1983.

[2] J. Burrell, "How the machine "thinks": Understanding opacity in machine learning algorithms," *Big Data & Society*, vol. 3, no. 1, 205395171562251–12, 2016.

[3] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[4] F. Pasquale, *The Black Box Society: The Secret Algorithms that Control Money and Information.* Cambridge, Massachusetts; London, England: Harvard University Press, 2015.

[5] G. Ciatto, M. I. Schumacher, A. Omicini, and D. Calvaresi, "Agent-based explanations in AI: Towards an abstract framework," In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*. Springer, Cham, 2020, pp. 3–20.

[6] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling, "Explainable agents and robots: Results from a systematic literature review," *Proceedings of the International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents & Multiagent Systems, 2019, pp. 1078–1088.

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

[8] D. M. Bailer-Jones, and C. A. Bailer-Jones, "Modeling data: Analogies in neural networks, simulated annealing and genetic algorithms," in *Model-Based Reasoning: Science, technology, values,* L. Magnani and N. Nersessian, eds. New York: Kluwer Academic/Plenum Publishers, 2002, pp. 147–165.

[9] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An approach to evaluating interpretability of machine learning," arXiv preprint arXiv:1806.00069, 2018.

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

[11] A. Rosenfeld, and A. Richardson, "Explainability in human–agent systems," *Autonomous Agents and Multi-Agent Systems*, vol. 33, no. 6, 673–705, 2019.

[12] A. Garfinkel, "Forms of explanation: Rethinking the questions in social theory," New Haven/London: Yale University Press, 1982.

[13] H. Felzmann, E. F. Villaronga, C. Lutz, and A. Tamò-Larrieux, "Transparency you can trust: Transparency requirements for artificial

intelligence between legal norms and contextual concerns," *Big Data & Society*, vol. 6, no. 1, 2053951719860542.

[14] A. Preece, D. Harborne, D. Braines, R. Tomsett, and S. Chakraborty, "Stakeholders in explainable AI," *arXiv preprint arXiv:*1810.00184, 2018.

[15] G. Ras, M. van Gerven, and P. Haselager, "Explanation methods in deep learning: Users, values, concerns and challenges," in *Explainable and Interpretable Models in Computer Vision and Machine Learning*, H. J. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, M. van Gerven, eds. Springer, Cham, 2018, pp. 19–36.

[16] R. Tomsett, D. Braines, D. Harborne, A. Preece, S. Chakraborty, "Interpretable to whom? A role-based model for analyzing interpretable machine learning systems," in *ICML Workshop on Human Interpretability in Machine Learning*. Stockholm, Sweden, 2018.

[17] F. Mairesse, and M. A. Walker, "Towards personality-based user adaptation: psychologically informed stylistic language generation," *User Modeling and User-Adapted Interaction*, vol. 20, no. 3, 227-278, 2010.

[18] U. Ehsan, P. Tambwekar, L. Chan, B. Harrison, and M. O. Riedl, "Automated rationale generation: a technique for explainable AI and its effects on human perceptions," in *Proceedings of the 24th International Conference on Intelligent User Interfaces.* ACM, 2019, pp. 263–274.

[19] M. F. Schober, and S. E. Brennan, "Processes of interactive spoken discourse: The role of the partner," in *Handbook of discourse processes,* A. C. Graesser, M. A. Gernsbacher, S. R. Goldman, eds. Mahway, New Jersey/London: Lawrence Erlbaum Associates, 2003, pp. 128–169.

[20] H. H. Clark, and M. A. Krych, "Speaking while monitoring addressees for understanding," *Journal of Memory and Language*, vol. 50, no. 1, 62–81, 2004.

[21] W. Wahlster, and A. Kobsa, "User models in dialog systems," in *User models in dialog systems*, A. Kobsa, W. Wahlster, eds. Berlin, Heidelberg: Springer, 1989, pp. 4–34.

[22] A. Cawsey, *Explanation and interaction: The computer generation of explanatory dialogues*. Cambridge M.A.: MIT Press, 1992.

[23] B. Sodian, "Theory of mind in infancy," *Child Development Perspectives*, vol. 5, no. 1, 39–43, 2011.

[24] R. Fusaroli, J. Raczaszek-Leonardi, and K. Tylén. Dialog as interpersonal synergy. *New Ideas in Psychology*, vol. 32, 147–157, 2014.

[25] H. H. Clark, *Using language.* Cambridge: Cambridge University Press, 1996.

[26] N. Sebanz, H. Bekkering, and G. Knoblich, "Joint action: bodies and minds moving together," *Trends in Cognitive Sciences*, vol. 10, no. 2, 70-76, 2006.

[27] C. Michel, C. Wronski, S. Pauen, M. M. Daum, and S. Hoehl, "Infants' object processing is guided specifically by social cues," *Neuropsychologia,* vol. 126, 54–61, 2017.

[28] M. J. Pickering, and S. Garrod, Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, vol. 27, no. 2, 169–190, 2004.

[29] D. Reitter, and J. D. Moore, "Alignment and task success in spoken dialogue," *Journal of Memory and Language*, vol. 76, 29–46, 2014.

[30] Fischer, K. *Designing Speech for a Recipient: The roles of partner modeling, alignment and feedback in so-called'simplified registers,* (Vol. 270). Amsterdam/Philadelphia: John Benjamins Publishing Company, 2016.

[31] B. F. Malle, "How people explain behavior: A new theoretical framework," *Personality and Social Psychology Review*, vol. 3, no. 1, 23–48, 1999.

[32] F. Keil, "Explanation and Understanding," *Annual Review of Psychology*, vol. 57, 227–254, 2006.

[33] H. Kotthoff, "Erklärende Aktivitätstypen in Alltags- und Unterrichtskontexten," in *Erklären im Kontext. Neue Perspektiven aus der Gesprächs- und Unterrichtsforschung*, J. Spreckels, ed. Hohengehren: Schneider, 2009, pp. 120–146.

[34] K. Ehlich, „Erklären verstehen—Erklären und Verstehen," in *Erklären. Gesprächsanalytische und fachdidaktische Perspektiven*, R. Vogt, ed. Tübingen: Stauffenburg, 2008, pp. 11–24.

[35] M. Morek, *Kinder erklären: Interaktionen in Familie und Unterricht im Vergleich.* Tübingen: Stauffenburg-Verlag, 2012.

[36] A. C. Graesser, N. K. Person, and J. P. Magliano, "Collaborative dialogue patterns in naturalistic one-to-one tutoring," *Applied Cognitive Psychology*, vol. 9, 495–522, 1995.

[37] S. M. Lundberg, and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, 4765–4774, 2017.

[38] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," *Journal of Computational and Graphical Statistics,* vol. 24, no. 1, 44–65, 2015.

[39] M. Ribeiro, T. Sameer Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining,* ACM, 2016, pp. 1135–1144.

[40] C. Molnar, *Interpretable machine learning.* Lulu. com, 2019.

[41] A.-C. Ngonga Ngomo, Di. Moussallem, and L. Bühmann. „A holistic natural language generation framework for the semantic web," in *Proceedings of Recent Advances in Natural Language Processing*, 2019, pp. 819–828.

[42] K. Sokol, and P. A. Flach, "Glass-Box: Explaining AI decisions with counterfactual statements through conversation with a voice-enabled virtual assistant," in *IJCAI*, 2018, pp. 5868–5870.

[43] D. H. Park, L. A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach, "Multimodal explanations: Justifying decisions and pointing to the evidence," *arXiv:*1802.08129, 2018.

[44] C. M. Mills, K. R. Sands, S. P. Rowles, and I. L. Campbell, "I want to know more!: Children are sensitive to explanation quality when exploring new information," *Cognitive Science*, vol. 43 no. 1, e12706, 2019.

[45] M. Morek, V. Heller, and U. Quasthoff, "Erklären und Argumentieren. Modellierungen und empirische Befunde zu Strukturen und Varianzen," in *Begründen—Erklären—Argumentieren. Konzepte und Modellierungen in der Angewandten Linguistik*, I. Meißner, and E. L. Wyss, eds. Tübingen: Stauffenburg Verlag, 2017, pp. 11–46

[46] N. Miyake, "Constructive interaction and the iterative process of understanding," *Cognitive Science*, vol. 10, 151–177, 1986.

[47] R. D. Roscoe, and M. T. H. Chi, "Tutor learning: the role of explaining and responding to questions," *Instructional Science,* vol. 36, no. 4, 321–350, 2007.

[48] P. Auer, "Introdction: John Gumperz' approach to contextualization," in *The contextualization of language*, P. Auer, and A. Di Luzio, eds. Amsterdam/Philadelphia: John Benjamins Publishing, 1992, pp. 1–38.

[49] P. Auer, and A. Di Luzio, eds. *The contextualization of language.* Amsterdam/Philadelphia: John Benjamins Publishing, 1992.

[50] H. H. Clark, and S. E. Brennan, "Grounding in communication." in Perspectives on socially shared cognition, L. B. Resnick, J. M. Levine, and S. D. Teasley, eds. American Psychological Association, 222–233, 1991.

[51] H. Buschmeier, *Attentive speaking. From listener feedback to interactive adaptation,* PhD thesis, Bielefeld University, 2018.

[52] M. Stone, and A. Lascarides, "Coherence and rationality in grounding," in *Proceedings of the 14th Workshop on the Semantics and Pragmatics of Dialogue*. Poznań, Poland, 2010, pp. 51–58.

[53] D. R. Traum, and J. F. Allen, "Discourse obligations in dialogue processing." in *Proceedings of the Thirty-third Annual Meeting of the Association for Computational Linguistics*, New Mexico, 1994, pp. 1–8.

[54] Visser, D. Traum, D. DeVault, and R. op den Akker, "A model for incremental grounding in spoken dialogue systems," *Journal on Multimodal User Interfaces*, vol. 8, no. 1, 61–73, 2014.

[55] D. Brown-Schmidt, S. O. Yoon, and R. A. Ryskin, "People as contexts in conversation," *Psychology of learning and motivation*, vol. 62, 59–99. 2015.

[56] J. De Ridder, "Mechanistic artefact explanation," *Studies in History and Philosophy of Science Part A*, vol. 37, no. 1, 81–96, 2006.

[57] C. Schulte, "Duality reconstruction–teaching digital artifacts from a socio-technical perspective," in *International Conference on Informatics in Secondary Schools-Evolution and Perspectives*. Springer, Berlin, Heidelberg. 2008, pp. 110–121

[58] I. Rahwan, M. Cebrian, N. Obradovich, J. Bongard, J. F. Bonnefon, C. Breazeal, J. W. Crandall, N. A. Christakis, I. D. Couzin, M. O. Jackson, N. R. Jennings, E. Kamar, I. M. Kloumann, H. Larochelle, D. Lazer, R. Mc Elreath, A. Mislove, D. C. Parkes, A. Pentland, M. E. Roberts, A. Shariff, J. B. Tennenbaum, and M. Wellman, "Machine behaviour," Nature, vol. 568, no. 7753, 477–486, 2019.

[59] M. J. Pickering, and S. Garrod, "An integrated theory of language production and comprehension," Behavioral and Brain Sciences, vol. 36, no. 4, 329–347, 2013.

[60] H. Buschmeier, and S. Kopp, Communicative listener feedback in human-agent interaction: Artificial speakers need to be attentive and adaptive," in *Proceedings of the 17th International Conference on Autonomous*

Agents and MultiAgent Systems. International Foundation for Autonomous Agents and Multiagent Systems, 2018, pp. 1213–1221.

[61] M. T. H. Chi, S. A. Siler, and H. Jeong, T. Yamauchi, and R. G. Hausmann, "Learning from human tutoring," Cognitive Science, vol. 22, 471–533, 2001.

[62] M. T. H. Chi, S. A. Siler, and H. Jeong, "Can tutors monitor students' understanding accurately?" Cognition and Instruction, vol. 22, no. 3, 363–387, 2004.

[63] K. Pitsch, A. L. Vollmer, K. J. Rohlfing, J. Fritsch, and B. Wrede, "Tutoring in adult-child interaction: On the loop of the tutor's action modification and the recipient's gaze," Interaction Studies, vol 15, no. 1, 55–98, 2014.

[64] K. Fischer, K. Foth, K. J. Rohlfing, and B. Wrede, Mindful tutors: Linguistic choice and action demonstration in speech to infants and a simulated robot," *Interaction Studies*, vol. 12, no. 1, 134–161, 2011.

[65] B. Wrede, S. Kopp, K. Rohlfing, M. Lohse, and C. Muhl, „Appropriate feedback in asymmetric interactions," Journal of Pragmatics, vol. 42, no. 9, 2369–2384, 2010.

[66] D. Wood, J. S. Bruner, and G. Ross, "The role of tutoring in problem solving," Journal of Child Psychology and Psychiatry, vol. 17, no. 2, 89–100, 1976.

[67] J. J. Gumperz, "Contextualization revisited," in *The contextualization of language*, P. Auer, and A. Di Luzio, eds. Amsterdam/Philadelphia: John Benjamins Publishing, 1992, pp. 39–53.

[68] I. Howley, T. Kanda, K. Hayashi, and C. Rosé, "Effects of social presence and social role on help-seeking and learning," in *Proceedings of the ACM/IEEE international conference on Human-robot interaction*, ACM, 415–422, 2014.

[69] A. Cangelosi, G. Metta, G. Sagerer, S. Nolfi, C. L. Nehaniv, K. Fischer, J. Tani, B. Belpaeme, G. Sandini, L. Fadiga, B. Wrede, K. Rohlfing, E. Tuci, K. Dautenhahn, J. Saunders, and A. Zeschel, "Integration of action and language knowledge: A roadmap for developmental robotics," IEEE Transactions on Autonomous Mental Development, vol. 2, no. 3, pp. 167–195, 2010.

[70] R. Mermelshtine, "Parent-child learning interactions: A review of the litera- ture on scaffolding," *British Journal of Educational Psychology*, vol. 87, 241–254, 2017.

[71] M. Amerian, and F. Mehri, "Scaffolding in sociocultural theory: Definition, steps, features, conditions, tools, and effective consideration," *Scientific Journal of Review*, vol .3, 756–765, 2014.

[72] K. J. Rohlfing, J., Fritsch, B. Wrede, and T. Jungmann, "How can multimodal cues from child-directed interaction reduce learning complexity in robots?" *Advanced Robotics*, vol. 20, no. 10, 1183–1199, 2006.

[73] J. S. Herberg, M. M. Saylor, P. Ratanaswasd, D. T. Levin, and D. M. Wilkes, "Audience-contingent variation in action demonstrations for humans and computers," *Cognitive Science*, vol. 32, no. 6, 1003–1020, 2008.

[74] H. M. Buhl, "Partner orientation and speaker's knowledge as conflicting parameters in language production," Journal of Psycholinguistic Research, vol. 30, no. 6, 549–567, 2001.

[75] A. L. Vollmer, K. Pitsch, K. S. Lohan, J. Fritsch, K. J. Rohlfing, and B. Wrede, "Developing feedback: How children of different age contribute to a tutoring interaction with adults," in *Proceedings of the IEEE 9th International Conference on Development and Learning*, 2010, pp. 76–81.

[76] G. Psathas, ""Talk and social structure" and "studies of work"," *Human Studies*, vol. 18, no. 2–3, 139-155, 1995.

[77] U. Quasthoff, V. Heller, and M. Morek, "On the sequential organization and genre-orientation of discourse units in interaction: An analytic framework," *Discourse Studies*, vol. 19, no. 1, 84–110, 2017.

[78] 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, 470, 2016.

[79] Selting, M., „Praktiken des Sprechens und Interagierens im Gespräch aus der Sicht von Konversationsanalyse und Interaktionaler Linguistik," in *Sprachliche und kommunikative Praktiken*, A. Deppermann, H. Feilke, and A. Linke, eds. Berlin: de Gruyter, 2016, pp. 27–56.

[80] A. Deppermann, H. Feilke, and A. Linke, Sprachliche und kommunikative Praktiken. Berlin: de Gruyter, 2016.

[81] H. Moore, C. Jasper, and A. Gillespie, "Moving between frames: The basis of the stable and dialogical self," Culture and Psychology, vol. 17, no. 4, 510–519.

[82] R. A. Wilson, and F. C. Keil, „The shadows and shallows of explanation," in *Explanation and Cognition*, F. C. Keil, and R. A. Wilson, eds. Cambridge M.A.: MIT Press, 2000, pp. 87–114.

[83] W. Imo, "Im Zweifel für den Zweifel: Praktiken des Zweifelns," in *Sprachliche und kommunikative Praktiken*, A. Deppermann, H. Feilke, and A. Linke, eds. Berlin, Boston: de Gruyter, pp. 153–176, 2016.

[84] G. Gonos, "'Situation' versus 'Frame': The 'Interactionist' and the 'Structuralist' analyses of everyday life," *American Sociological Review*, vol. 42, 854–867, 1977.

[85] D. Tannen, and C. Wallat, "Interactive frames and knowledge schemas in interaction: Examples from a medical examination/interview," *Social Psychology Quarterly*, vol. 50, no. 2, 205–216, 1987.

[86] T. Matzner, P. K. Masur, C. Ochs, and T. von Pape, "Do-It-Yourself Data Protection—Empowerment or Burden?" in *Data Protection on the Move,* S. Gutwirth, R. Leenes, and P. De Hert, eds. Dordrecht: Springer, 2016, pp. 277–305.

[87] D. Neyland, "Bearing Account-able Witness to the Ethical Algorithmic System," *Science, Technology & Human Values*, vol. 41, no. 1, 50–76, 2016

[88] A. McStay, "I consent: An analysis of the Cookie Directive and its implications for UK behavioral advertising," *New Media & Society*, vol. 15, no. 4, 596–611, 2012.

[89] E. Braf, "Knowledge or information," in *Proceedings of the IFIP TC8 / WG8.1 Working Conference on Organizational Semiotics: Evolving a Science of Information Systems*. Kluwer, B.V., NLD, 2001, pp. 71–90.