Secondary Publication

Schmid, Ute; Wrede, Britta

What is Missing in XAI So Far? : An Interdisciplinary Perspective

Date of secondary publication: 13.11.2023

Version of Record (Published Version), Article

Persistent identifier: urn:nbn:de:bvb:473-irb-916432

Primary publication

Schmid, Ute; Wrede, Britta (2022): „What is Missing in XAI So Far? : An Interdisciplinary Perspective“. In: Künstliche Intelligenz : KI ; Forschung, Entwicklung, Erfahrungen ; Organ des Fachbereichs 1 Künstliche Intelligenz der Gesellschaft für Informatik e.V., GI, Vol. 36, Nr. 3-4, pp. 303-315, Berlin: Springer, doi: 10.1007/s13218-022-00786-2.

Legal Notice

This work is protected by copyright and/or the indication of a licence. You are free to use this work in any way permitted by the copyright and/or the licence that applies to your usage. For other uses, you must obtain permission from the rights-holder(s).

This document is made available under a Creative Commons license.

The license information is available online:
https://creativecommons.org/licenses/by/4.0/legalcode

KI - Künstliche Intelligenz (2022) 36:303–315
https://doi.org/10.1007/s13218-022-00786-2

TECHNICAL CONTRIBUTION

What is Missing in XAI So Far?

An Interdisciplinary Perspective

Ute Schmid[1] · Britta Wrede[2]

Received: 28 October 2022 / Accepted: 21 November 2022 / Published online: 7 December 2022
© The Author(s) 2022

Abstract
With the perspective on applications of AI-technology, especially data intensive deep learning approaches, the need for methods to control and understand such models has been recognized and gave rise to a new research domain labeled explainable artificial intelligence (XAI). In this overview paper we give an interim appraisal of what has been achieved so far and where there are still gaps in the research. We take an interdisciplinary perspective to identify challenges on XAI research and point to open questions with respect to the quality of the explanations regarding faithfulness and consistency of explanations. On the other hand we see a need regarding the interaction between XAI and user to allow for adaptability to specific information needs and explanatory dialog for informed decision making as well as the possibility to correct models and explanations by interaction. This endeavor requires an integrated interdisciplinary perspective and rigorous approaches to empirical evaluation based on psychological, linguistic and even sociological theories.

Keywords Explainable AI · Human-centred AI · Hybrid AI

Mathematics Subject Classification 68T01: general · 68T05: learning and adaptive systems

1 Introduction

With the growing number of real-world applications of AI-technology, the question about how to control such systems has become a growing concern [73]. Especially the rise of models with large degrees of freedom and relatively little pre-defined structure in the last decade has increased need for understanding learned models through explanation. The requirement that humans need to be able to understand on what information an AI-system relies to derive a specific output, has been identified as crucial early in the discussion and had been labeled explainable AI (XAI) [2].

Originally, the focus has been mostly on visual explanations for image classification [67, 74]. At first, such explanations by visual highlighting have been – rather naively – seen as a suitable means to make decisions of a black box system transparent to end-users and thereby inspire trust. However, this view has been pointed out as being simplistic from early on from an interdisciplinary perspective [56]. Consequently, XAI methods have been developed with different stakeholders in mind. Visual highlighting based on relevance of information in the input has been shown to be especially helpful from a developer's perspective [61] with the goal to identify overfitting of models [52] (see Fig. 1). As a second stakeholder group, end users are considered. Here, the European Union's General Data Protection Regulation (GDPR) boosted research initiatives that target XAI as a means to support the right for explanations – especially in Europe [89].

Furthermore, it has been noted that especially critical application areas of AI such as medicine, industrial production, or justice also require transparency and tractability

✉ Britta Wrede
bwrede3@uni-bielefeld.de

Ute Schmid
ute.schmid@uni-bamberg.de

1 Cognitive Systems, University of Bamberg, Bamberg, Germany

2 Systems Engineering for Cognitive Robotics and Cognitive Systems, University of Bremen and Center for Cognitive Interaction Technology (CITEC), Bielefeld University, Bielefeld, Germany

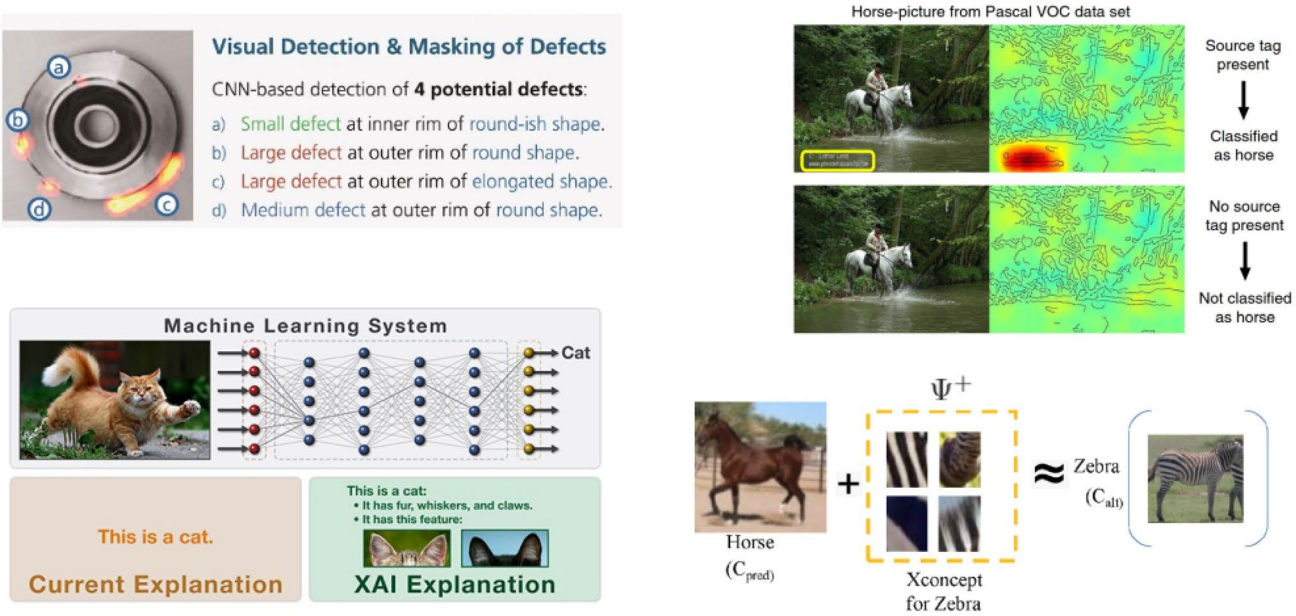

Fig. 1 Different types of explanations: visual highlighting to detect overfitting [52] (top right), combining visual and verbal explanations [60] (top left), explaining verbally and with visual prototypes [33] (bottom left), explaining by near miss contrastive examples [4] (bottom right)

which can serve as justification and to provide trustworthy support by enabling domain experts to comprehend how the AI system has come to a specific proposition and which part of the input data has contributed most to a decision [40, 60]. For domain experts, explanations have mostly the function to support decision making. In this context, there are high demands on the faithfulness of explanations and on their expressiveness [77]. Besides system developers, end-users, and domain experts, recently additional stakeholder groups have come into focus such as affected parties, deployers, and regulators [50].

While the term "explainable AI" has found its way into the wider public discourse, the problem itself has generated a range of different terms that have been used in addition to explainability [40] – namely comprehensibility [59] and interpretability [23, 58]. These latter two terms have been mainly used with respect to the transparency of machine learned models. Nowadays, the term interpretable machine learning mostly refers to symbolic, white-box machine learning approaches which can be used either as surrogate models [20, 21, 65] or as alternative for black-box models derived from training deep neural networks or other statistics approaches [49, 72]. The choice of the term "explainability" gave rise to some misunderstandings outside the XAI research community, being for example interpreted such that some expert explains what is meant by artificial intelligence or how a specific system works. Consequently, sometimes now "explanatory" is used instead [3, 47, 83]. Furthermore, the characteristics of an AI-system to provide an explanation for a decision has been proposed a crucial feature of so-called third wave of AI systems[1], superseding the second wave of AI focussing on data-intensive blackbox machine learning approaches, which have followed to first wave of AI knowledge-based symbolic approaches [41, 79].

The topic of explanations has been addressed in AI research already long before the current interest. In the context of expert-system research, approaches to generate verbal explanations from reasoning traces have been proposed [18]. In the German Journal of Artificial Intelligence (KI – Künstliche Intelligenz) the topic of explanations has been addressed in the context of expert systems [45] and has been dedicated a special issue in 2008[2] presenting research on explanations in different areas of AI such as case-based reasoning and theorem proving [70]. More recently, explanations were researched extensively in the context of recommender systems [86] and in interactive systems [14, 31, 47]. In addition to visual explanations and other methods to highlight feature relevance, verbal explanations as proposed in symbolic AI research receive new interest, especially for explanations intended for domain experts in complex relational domains [60, 80, 91, 92] (see Fig. 1). Furthermore, explanations by example are investigated [37, 43].

Over the last years, on the one hand, a large variety of computational approaches to XAI have been developed, and, on the other hand, the topic of explanations has received

[1] http://hlc.doc.ic.ac.uk/3AI_HLC_2019.html. https://hessian.ai/projects/3ai-the-third-wave-of-ai.

[2] https://dblp.uni-trier.de/db/journals/ki/ki22.html

more attention in different disciplines such as human-computer interaction, educational and cognitive psychology, cognitive science as well as sociology and ethics. Furthermore, the topic has been treated in numerous surveys [2, 23, 34, 87]. Nevertheless, it is our impression that rigorous integration of XAI methods and insights from different disciplines is still missing. In the following, we will first introduce building blocks of an interdisciplinary perspective on XAI. Afterwards, we highlight selected aspects which, in our opinion, should receive more attention in future XAI research: faithfulness and consistency of explanations, adaptability to specific information needs and explanatory dialog for informed decision making, the possibility to correct models and explanations by interaction, as well as the need for an integrated interdisciplinary perspective and rigorous approaches to empirical evaluation based on psychological theories.

2 An Interdisciplinary Motivation

The AI community has developed a plethora of approaches for making AI decisions or models better understandable mostly to developers and AI experts. Only very recently the question what kind of explanations lay users actually need in their everyday interactions with AI and how their cognitive system affects their processing and ultimately their understanding of explanations has come into focus and points towards more human-centered approaches of explaining. We therefore argue to strengthen research in those XAI areas that are supported by interdisciplinary research with the aim to take the human explainee – the addressee in an explanation setting – and the situational context of the application into account.

In this paper we especially focus on the following perspectives:

Philosophy of science. Explanations have been studied for a long time in the field of philosophy of science addressing the question what are good scientific explanations. According to this explanations should follow logical rules in a deductive way [36] which is very similar to the road XAI has followed in its beginnings and is still pursuing. However, more recently it has been noted that lay persons do not follow such strict rules in their everyday explanations [42] and the understanding thereof. Keil argues that, for example, people are often not aware of why they prefer one explanation over another, thus apparently using implicit and not necessarily logical strategies to evaluate explanations. Also, they tend to construe causal explanations based on very sparse data, often simply on single-trial learning. For humans, the quality of an explanation is not tied to its logical correctness and completeness. Rather, they judge explanations along the dimensions circularity, relevance and coherence. Yet, it has been shown that circularity can be difficult for humans to detect [68]. Although understanding explanations is important for humans in order to carry out (complex) tasks and to achieve their goals, humans are often not aware that they have not understood an explanation, a phenomenon which has been termed "the illusion of explanatory depth" [71]. Thus, everyday explanations as they are used by most people in their daily lives seem to follow very different mechanisms than scientific ones.

Cognitive Psychology. In general, humans seldom make sense of their environment by explicit computation and evaluation of different hypotheses. Rather, they rely on heuristics that have proven helpful and mostly true in many situations and have often found their way into the human cortex. These heuristics can actually be interpreted as biases that shape how humans make sense of their complex environment and interactions such as explanations. These biases are well known in the literature. For example, [91] refer to cognitive biases in medical decision support systems. One frequent bias that unexperienced physicians experience is the positivity bias, i.e. once the physician has formed a hypothesis s/he searches for confirmation of this hypothesis instead of searching for falsification which would provide a much higher information gain. The anchoring bias is the tendency to rely too heavily on the first piece of information when making decisions. The availability bias is the tendency to overestimate the likelihood of events with greater "availability" in memory. This greater availability in memory can be achieved through more recent memories, e.g. a very rare disease which has been recently encountered, or through unusual or emotionally loaded events. Another bias in this context is a cognitive dissonance reduction bias, such as effort justification which is the tendency to attribute greater value to an outcome if one had to put effort into achieving it. In everyday situations we continuously encounter biases through prejudices and stereotypes which we also apply to explaining situations involving AI systems. Especially in the context of decision support systems Berkson's paradox can be difficult to overcome. It is the tendency to misinterpret statistics involving conditional probabilities which are fundamentally important in order to understand decision support suggestions from an AI system in a given context.

For XAI applications this means that the explaining system should actively search for cues that indicate that the human explainee is following heuristics and cognitive biases rather than logically correct reasoning and provide active support against these [90].

Educational Psychology. In the context of teaching explaining plays an important role. However, in this context it has been observed that the behavior or engagement of the learner seems to be at least equally important. It should be noted, that this kind of research focuses on scientific explanations rather than everyday explanations. However, we

believe that its focus on engagement is a valuable perspective also for XAI.

[16] have proposed a model that targets the involvement of the explainee in order to enable her to ask questions and actively generate new insights through reasoning and discussing. In detail, the ICAP model (for Interactive, Creative, Active, and Passive engagement behavior) [17] targets teaching situations and postulates that different learner engagements yield different learning outcomes. More specifically, according to this account the least learning success is achieved by passive engagement alone, i.e. attentive listening or reading. Active engagement which requires manipulative behaviors such as repeating and rehearsing as achieved e.g. by watching explanations videos yields better learning outcomes. Constructive engagement which is achieved by generative behaviors such as "reflecting out loud and self-explaining" which require to actively draw inferences based on one's mental model achieves better results. Best results are postulated by co-generative collaborative behaviors which are characterised by dialogical interactions involving "defending and arguing a position, and discussing similarities and differences" which challenge the learner to co-infer together with a peer or the teacher new propositions. Indeed, many studies support the basic assumptions of the ICAP model in lab and teaching contexts (e.g. [19, 30]). Insights from education, thus, strongly suggest that explanations should be embedded in a dialogical interaction allowing the explainee to be actively engaged and challenge those aspects of an explanation that have not been understood.

Developmental Linguistics. Explanations and teaching play an important role in parent-child interactions and important concepts have been developed in this context.

Scaffolding is a frequent and highly useful strategy that parents use for teaching their children [11, 93]. The specific benefit of scaffolding has been identified as enabling the child - or explainee - to solve a problem which would be beyond his/her current capabilities. It means that the teacher is carrying out those elements that the learner can not yet do. In a joint task setting with children this often pertains to motor actions, but it can also mean to explicate a reasoning step in a more abstract explanation situation or to ask a (suggestive) question, enabling the explainee to complete those elements of the problem that are within his/her range of competence and to achieve a successful completion of the task. A feature that so far has received less attention pertains to the emotional quality of scaffolding. In general, scaffolding interactions have a highly positive emotional component that serves as a strong motivator to the learner. However, how these emotional processes work in detail together with the understanding process remains largely unknown.

Taking features of scaffolding into account, as provided by [93], scaffolding requires from the explainer to provide contingent feedback on the learner's actions, to give hints or tips, to provide instructions and explanations, and to ask questions. Importantly, the teacher should introduce and divide the task into manageable pieces to allow for continuous progress in dealing with the challenge. This entails to provide meaningful associations between objects, concepts and actions and to make use of non-verbal means.

Scaffolding, thus, requires the ability to monitor the learner's attention and understanding process in order to identify knowledge gaps that need do be addressed by the teacher.

New perspectives on explaining in XAI. Humans' explanations are characterised by an interactive account which follows specific sequences [44, 63]. Such interactive explanations enable adaptation to the explainee's level of understanding and should be the basis of any explanation approach [69].

More specifically [69] challenge a range of often implicitly made assumptions that dominate much of current XAI research. According to this, XAI targets scientific explanations for which accuracy and completeness are mandatory. Whereas in their view "everyday explanations can be understood as partial descriptions of causes that enhance understanding in the user in terms of why something happened". This places the user's understanding in the center of the explaining process rather than the explanandum, i.e. the elements that need to be explained. Furthermore, [69] argue that personalisation which is a standard approach in Human-Computer Interaction and also being explored in XAI is not an adequate strategy to adapt to individual differences as the understanding process is dynamic, that is the level of understanding is continuously changing and adaptation, thus, has to take place at each step. This also entails to model the explainee's prior knowledge which can also vary widely. To address this, a conceptual framework is proposed that allows to "co-construct understanding" through a "bidirectional interaction involving both partners in constructing the task and its relevant (recipient-oriented) aspects". This idea of co-construction seems to carry interesting similarities to the above mentioned concept of "co-generation" from the ICAP model which has been shown to yield better learning results. This co-construction requires that the explainer closely monitors the explainee for signs of (non-)understanding and to provide scaffolding where understanding becomes difficult (cf. Fig. 2). This also requires the explainer to continuously update her partner model M_{ER}. Through this contingent process of monitoring and scaffolding the explanans is constructed, i.e. the co-constructed explanation which only encompasses a part of the explanandum and that (optimally) addresses those aspects that are relevant to the explainee.

Mental Models, Partner Models and Theory of Mind. Thus, an important prerequisite for construing as well as understanding explanations are mental models and cognitive processes that allow to represent, retrieve and apply relevant

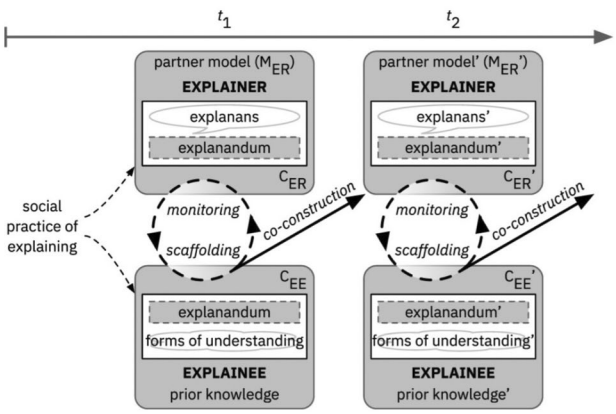

Fig. 2 Explanation process according to [69] yielding a co-construction of the explanation by the active contributions of the explainer (ER) and the explainee (EE) by continuously monitoring and scaffolding one another. It requires a mental model by both, the explainer as well as the explainee

aspects of a phenomenon [42] as well as a model of the partner and the partner's model.

Human explainers utilize a mental model about the explainee's understanding, i.e. they form hypotheses about what the explainee has understood and what aspects need to be corrected or explained again, or should be attended to more strongly. For example, [16] investigated how mental models are formed by students and how they are addressed by the teacher when explaining learning content. According to her research, explainers form general models about their tutee's understanding that represent canonical misunderstandings. However, they found surprisingly little evidence that explainers also were able to build models of individualized misconceptions that go beyond canonical misunderstandings [16]. Nevertheless, having a model of the explainee's understanding seems to be important for successful explanation processes. Indeed, it could be shown that an Intelligent Tutoring System with contingent and fine-grained tutoring sub-steps achieved a similar learning effect as a human tutor [88].

In HCI and HRI partner models are used that represent relevant aspects of the ongoing interaction and the partner's knowledge in terms of perception, understanding, acceptance and agreement [13]. In order to incorporate the interaction partner's perspective into the own planning and reasoning Theory-of-Mind models [51] are being developed and investigated in various contexts. Such a mentalistic approach might provide a good basis for explaining in XAI. However, so far research on partner models or TOM models in the context of XAI seems to be very limited.

While many authors from the social sciences demand for user-driven approaches that are grounded in real world situations, there actually exists almost no empirical data on how AI is used in real life and what explanatory demands actually arise from these situations.

3 Assessment of Needs for XAI

Based on the observations above, in the following, we will highlight some aspects on XAI which are, in our opinion, crucial for successful real-world applications of AI in human-in-the loop settings.

3.1 Faithfulness and Consistency of Explanations

In order for an explanation of a recommendation or a behavior from an AI system to be helpful, it is crucial that it does not mislead the recipients of the explanation. This means that the explanation should approximate as precisely as possible which information (and possibly to what degree) the AI system has taken into account in order to generate its output. This property of an explanation is called faithfulness or fidelity [58, 94, 96].

Faithfulness of explanations is also not guaranteed when humans give causal explanations or justifications for their behavior [42]. However, for trustworthiness of AI systems, it is crucial that XAI methods provide explanations which reflect what information a model has used to generate an output. Fidelity is relevant for all types of explanations – for global explanations of the entire model as well as for local explanations related to individual inputs. In some cases, explanations can be used as surrogates for the originally learned models to classify new instances or make predictions. In this case, predictive accuracy has to be assessed not only for the model but also for the explanation. It should hold that the accuracy of the explanation is not (significantly) lower than that of the original model.

If different explainability approaches or variants of one approach result in different explanations – that is, explanations are inconsistent –, this is an indication for low fidelity. This is, for example, the case when LIME (Local Interpretable Model-agnostic Explanations) is applied with different superpixel methods [75]. LIME is one of the first XAI methods which have received a lot of attention [67]. As a model-agnostic explanation method, it does not interfere with the learned model. Instead, for a given input a set of perturbed variants are generated by hiding parts of the information. For the case of images, this is realized by clustering pixels into groups of superpixels which are hidden. For an explanation to be faithful, it should not be the case that highlighting what information the (untouched) model has used to come to a decision is influenced by the underlying clustering strategy.

Especially for black box machine learning, a faithful explanation is crucial to point out deficiencies of the underlying model, such as overfitting. This might lead to

overconfidence in a system based on the clever Hans effect – that is a model suggesting to a human that it accurately identifies some concept – be it a horse in a photo [52] or a tumor in microscopic image of tissue sample [10]. Research on methods to evaluate fidelity of explanations has only been started, currently, assessment mainly relies on perturbation methods [94].

3.2 Adaptability to Specific Information Needs

One approach to adapt explanations to the user's need is by specifying groups that are likely to have coherent questions with respect to the explanandum. In this vein, adaptation to groups such as domain experts vs novices or personalization [12, 46] have been suggested. Personalization is a process where typically the users themselves specify preferences or characteristics of themselves which are then used to parameterize the ensuing explanation with respect to different dimensions. It thus serves as a prior for the ensuing interaction which is, however, not updated by evidence for understanding or non-understanding by the user. [57] distinguish between AI-novices, Data Experts and AI Experts, thus focusing on expertise in AI systems vs in the domain where the data comes from, e.g. medicine with medical experts.

Some studies reveal the need for individually different interaction modes and parameters. For example, [55] found that users preferred different timing when blending in explanations based on hovering gaze of the user. Rather than preferences, some approaches target to determine the internal state of the explainee through observable features in order to estimate the success of different explanations and, thus, the comprehension of the user. One such approach is using facial expressions and it revealed that some features such as frowning – which is related to the activation of specific facial muscles – were related to a non-effective use of the explanations [32], i.e. to problems of understanding. Such results are promising with regard to future use of automatically detectable signals that can be used for monitoring and scaffolding.

To realize adaptability to specific types of content, it is prerequisite that different options how to explain something to someone are available. In a situation where it is crucial to get a fast response, visual highlighting of relevant information might be a good option. In cases, where decisions have a critical outcome such as in medical diagnostics, education, or law, different modalities might be useful to allow the human decision maker to better evaluate the proposition of the model and to gain better insight into relevant information on which the proposition is based. Verbal explanations are helpful to communicate complex relational dependencies, such as the location of tumor tissue relative to fat tissue [10, 77]. Furthermore, they might be given on different levels of detail: For interpretable models, this can be realized by following the reasoning trace from a set of rules [26, 76]. For deep neural network models, a similar strategy might be realized by extracting information from layers on different levels [15, 64].

Explaining by a counterfactual [89] or a contrastive example [22] helps to point out what is missing from a specific instance such that it would be classified differently. For structural data, near miss explanations can be constructed by identifying the minimal change in a rule resulting in a different class [66]. This principle has been introduced as alignment based reasoning in cognitive science and has been shown to be highly helpful to understand what the relevant aspects of a concept are [28]. For image data, different similarity measures have been explored to select either an instance which is prototypical for the class as represented in a learned model [43] or an example near the decision boundary of the model – either for the same class [43] or for a different class [37]

One perspective on adaptability is to automatically decide what kind of explanation is most helpful in a specific context. This has been, for instance, proposed in the context of companion systems which should be able to identify the emotional state and the intention of a user [9]. In the context of social robotics, it has been proposed to use reinforcement learning to generate user-adaptive behavior [35]. For such automated approaches, the models underlying explanation selection might be black boxes themselves such that the adaptive behavior of the system might require explanation itself. Alternatively, different options for explanations might be offered to the user who select what they find appropriate interactively [26, 47].

3.3 Explanatory Dialogs for Informed Decision Making

In human-human communication, explanations in different modalities and on different levels of detail are often provided in form of questions and answers. This is true for everyday communication, communication in the context of collaboration in work settings, and also in consultations between experts such as medical doctors or lawyers and patients or clients.

Mostly, an utterance or behavior of an interaction partner is just accepted without the need of an explanation. Sometimes, a justification is asked for, sometimes a causal explanation is requested. There are mostly two settings, in which decision making is based on more complex explanatory dialogs: If decisions are sensitive or involve serious risks, or in educational contexts. In the first case, the dialog might be between two experts, such as two medical decision makers or two quality engineers, or between an expert and a client. That is, dialog can be on one level of expertise

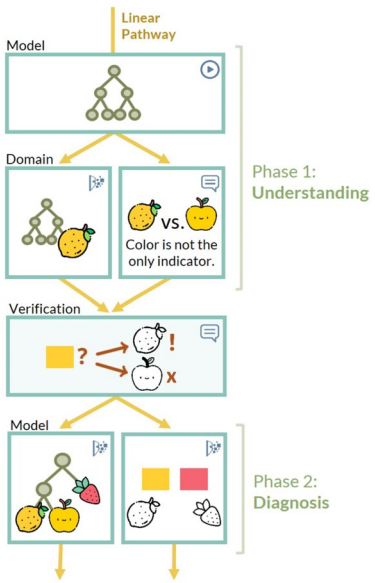

Fig. 3 Proposed structure of explanation process with multiple interaction types [24]. Phase 1 foresees an explanation phase which is followed by a verification phase. Only then the system allows the user to initiate interaction for achieving a diagnosis

from the recorded context or the human can ask explicitly for specific formats, such as *Please give me an example for a typical form of tumor class 3.* or *Can you be more specific how you define an intrusion.* [10].

Dialogical approaches to XAI follow diverse strategies. In [24] suggest a multi-stage dialog approach (cf. Fig. 3) that starts with a general explanation phase of the model and the domain ("Phase 1: Understanding") which is followed by a Verification Phase which needs to be mastered in order to proceed to the explanation of the decision at hand ("Phase 2: Diagnosis"). These phases can contain dialogical parts as well as videos, texts or visualizations. Phase 1 has the goal to provide the explainee with an initial knowledge base. Variations are foreseen with respect to the simplification strategy and the modality (visualization or text) that can be selected by the user.

In [26] propose an explanatory approach allowing rule-based and example-based explanations. They focus on explanations for image-based classification tasks in relational domains. Rule-based explanations are generated from a Prolog model [59], example-based explanations allow for selection of prototypes as well as near-miss examples [37]. An application example for medical diagnostics is given in Fig. 4.

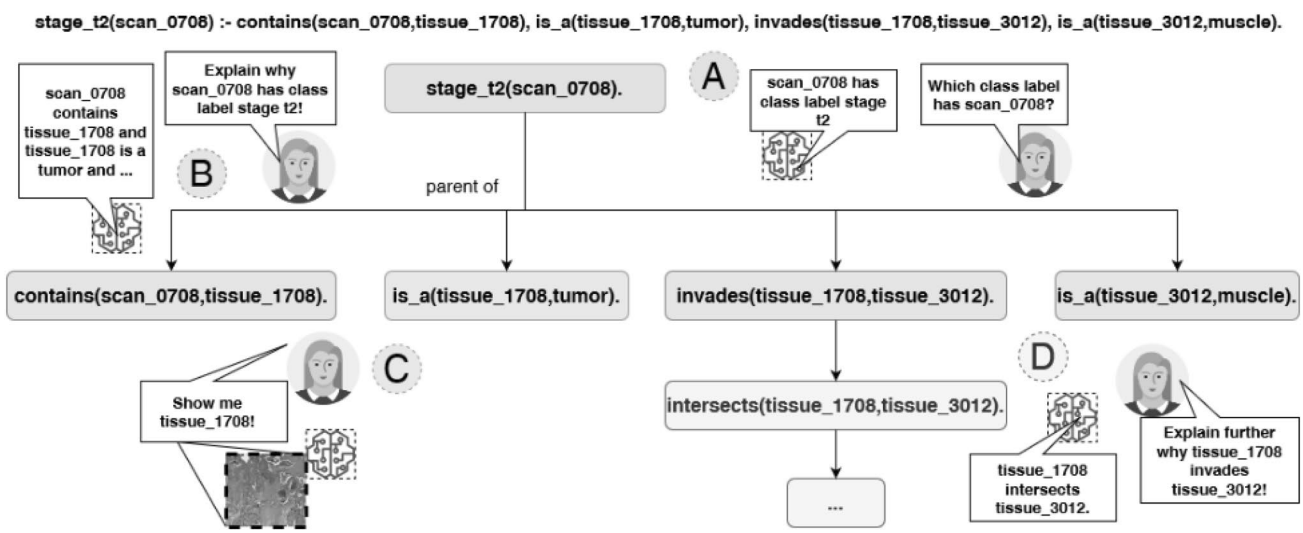

Fig. 4 An explanatory tree for `stage t2(scan 0708)`, that can be queried by the user to get a local explanation why scan 0708 is labeled as T2 (steps A and B). A dialogue is realized by further requests, either to get more visual explanations in terms of prototypes (step C) or to get more verbal explanations in a drill-down manner (step D)[27]

or, communication is asymmetrical between one expert or teacher and one layperson or student.

The same variants observed in human-human communication should be possible for the interaction between humans and AI-systems. That is, an explanatory dialog has to be adaptive with respect to the specific information needs. Again, the most helpful modality for an explanation and the adequate level of detail might be determined by the system

3.4 Interaction and Corrigibility for Human-in-the-Loop Decision Making

Approaches to interactive machine learning [82] have first been proposed in the context of query-based machine learning [6]. More recently, interactive learning has been considered mainly in the context of human-computer interaction

research [25, 38, 39]. The topic of explanations has not (explicitly) been considered in such approaches. The main goal of research has been to provide interfaces for effective interaction and correction of models such as feature selection or label corrections [8].

Interactive machine learning allows to inject knowledge into the process of model induction in form of specific corrections. However, in many applications, the underlying model is a black box. To make suitable corrections, especially in complex domains such as medical diagnostics, it is necessary or at least helpful, if the user first understands, based on what information the model has derived a specific decision [78, 83].

Explanations were introduced to interactive machine learning in form of explantory debugging to support adaptation for personalized assistants [47]. More recently, several approaches to combine XAI and interactive machine learning have been proposed [82].

Interactive accounts to provide local explanations, i.e. explanations of a system decision for one data point, are popular. For example, they allow the user to play around with different values to see how decisions would change for different values [90].

An approach which combines the model-agnostic explanation method LIME [67] with the possibility to correct the underlying machine learned model is CAIPI [83]. In LIME, the information which has been most relevant for the model to come to a specific classification decision is highlighted as explanation. CAIPI allows the user to provide a counterexample to correct a classification which has been right, but for the wrong reasons.

In a medical context an interactive approach to support physicians in diagnosing patients was developed in a user-centric way [90]. The resulting interface displaying key values of a patient together with the most likely "system predicted risks" and the associated feature attributions by vital values allowed the users to select one of the predicted risks to inspect counterfactual rules indicating key rules for each prediction. For example, the explanations suggest that the system has inferred a shock state because of low oxygen saturation and blood pressure of the patient (cf. Fig. 5).

Such interactive interfaces – especially when developed in collaboration with the envisioned system users – are an elegant solution to provide relevant information in an interactive way without actually assessing the user's level of understanding and reacting to it. In these approaches, the interaction is completely driven by the user.

When looking for inspiration how to model the user with respect to his/her level of understanding, it is worthwhile to take into account approaches proposed in the context of intelligent tutoring systems (ITS) [5, 62]. A typical ITS architecture has three components: A domain model which reflects the knowledge or skills to solve reasoning or problem solving tasks in a specific domain, a student model which identifies missing concepts or misconceptions of an individual learner, and a tutoring model which gives feedback in a way that addresses the specific information need of the learner to be able to solve a task correctly. The learner might be guided to a solution by a Socratic dialogue [29] or by example solutions for similar problems [95].

More recently, approaches have been proposed that not only target at increasing the learner's understanding but to motivate and "empower" them to ask specific questions that are especially helpful to get a better understanding of the topic at hand [1].

When explanations of AI systems are intended to support domain experts or end-users in decision making, the rich fund of ITS methods might be a source of inspiration [84].

3.5 Rigorous Empirical Evaluation

Despite that fact that the goal of XAI is to enable the user to understand why a system has come to a certain decision in order to (1) decide whether or not to trust this decision, (2) whether to correct the system, or (3) to possibly even derive new insights from it, evaluations often focus on predictive accuracy alone and omit assessment of the comprehensibility and usefulness of a system response for (a specific group of) users with a specific information need. An interesting approach to compare XAI approaches with respect to faithfulness is given by [7] which also investigates the agreement of the explanation with human rationales which are given through saliency scores for words. This second criterion for evaluating AI-systems has already been addressed by Donald Michie [54]. He proposed to extend evaluation of machine learning approaches from predictive accuracy to operative effectiveness. That is, that the output of the machine learning system increases the performance of a human to a level beyond that of the human studying the training data alone [59].

Both predictive performance as well as operative effectiveness need to be evaluated empirically. However, while the first criterion can be evaluated by calculating performance metrics for selected data sets, the second criterion must be evaluated by empirical studies involving humans.

In order to allow potential users, developers and scientists to get a first fast understanding of different explanation approaches [81] proposed so called *Explainability Fact Sheets* which are an attempt to provide a comprehensive description of the capabilities of an explaining framework such as LIME. While it mainly focuses on mathematical correctness and completeness issues through its *functional requirements* it also addresses *operational requirements* as well as user oriented aspects with the *usability requirements*. Safety is also explicitly addressed and information about

Fig. 5 Interactive user interface for medical doctors showing the raw features ("24 Hour Vitals") at the top with a visualisation of the predicted risks of different diagnoses at the top. Below this, an overview of the most relevant features for each of the different hypotheses is provided and the bottom part provides access to counterfactual rules explaining under what conditions the system would have selected a different hypothesis [90]

validations through user studies or synthetic experiments is also provided in the Fact Sheets.

A dedicated framework for evaluation has been proposed by [57] who foresee an approach that depends on the user group (cf. Fig. 6). According to their approach, XAI systems that target novices should be evaluated with respect to the user's mental model and the human-AI performance which is of course task dependent. Subjective evaluations should address user trust and reliance as well as usefulness and satisfaction. XAI systems for data experts, in contrast, should be evaluated based on task performance and model performance. AI experts who have a different agenda when dealing with their own created models allow to evaluate

	XAI Target Users		
	AI Novices	**Data Experts**	**AI Experts**
Design Goals	Transparency User Trust Bias Mitigation Privacy Awareness	Model Visualization and Inspection Model Tuning	Model Interpretability Model Debugging
Evaluation Measures	User Mental Model Usefulness and Satisfaction User Trust and Reliance Human-AI Performance	Task Performance Model Performance	Explainer Fidelity Model Trustworthiness

Fig. 6 Different design goals and evaluation measures for different target user groups as proposed by Mohseni at al. (adapted from [57])

more system oriented features such as explainer fidelity and model trustworthiness.

While the XAI research community has developed a large body of research on methods for explaining classifier behavior, dedicated user studies or participatory development approaches of XAI are comparatively scarce. This means that insights into what the relevant aspects of explaining AI to users are are still limited. In a review on user studies [48] identified five specific objectives and goals of XAI for end users: (1) understandability, (2) trustworthiness, (3) transparency, (4) controllability, and (5) fairness. They further identified a range of elements of an AI systems that require explanation or at least the ability for inspection. Accessibility to the raw data as well as information about the application and the situation are important to end users. Also, the model itself can require explanation. Regardless of the explanation target, the authors formulated general and specific design recommendations. According to these explanations need to be personalized, on demand and focused on key functionalities rather than addressing the complete system. One important insight concerns the recommendation to link explanations as much as possible to the mental models that the users have of the system. This is in line with other observations indicating that explanations need to be shaped towards the users' individual needs.

Besides usability requirements, *cognitive requirements* might be assessed and user studies might be complemented by rigorous experiments. For instance, the impact of different aspects such as simplicity or probability of an explanation on its perceived quality has been investigated in a series of experiments by [53]. The impact of verbal explanations in a joint human-AI decision making scenario has been assessed by contrasting performance of an experimental group receiving explanations with a control group [85]. Empirical research based on theories and methods of experimental cognitive psychology currently are scarce while there is growing number of user studies following the tradition of human factors research.

4 Conclusions and Outlook

Over the last years, research on XAI has broadened from focusing on developer oriented explanations of the AI model to consider other stakeholders such as end users, domain experts, affected parties, deployers or regulators. This has lead to the general insight that explanations need to be adaptive to individual user needs and take human cognitive processes into account that require certain strategies and may be subject to biases. This requires a strong interdisciplinary approach in order to equip mental and partner models with those features that are relevant for the contingent explaining process. Also, linguistic expertise is needed in order to develop dialog modeling approaches that can model contingent multi-modal interaction.

Accordingly, evaluations need to address multiple dimensions ranging from the correctness and performance of the classification and explanation model to the effect that the explanations generated by this system have on users in their specific contexts.

In our view, problematic issues of XAI that should receive more attention in the future arise from the pragmatics, that is, the specific use and context of an AI system in a concrete setting. This situatedness demands for a new perspective on evaluation and has led to a debate on the question under what conditions specific explanatory approaches are helpful and how this might be measured. For example, is an explanation helpful when the explainee is being supported (or "enabled") to take the "optimal" decision according to the decision support system? This entails further questions regarding among others effects of non-rational factors such as emotions or social biases.

A second strand of discourse pertains to the question of faithfulness. There is currently only little research on how faithful explainers actually are, that is, whether the proposed features are indeed the most relevant features that have lead to a certain decision, and how this can be measured reliably. This also impacts fairness as unfaithfulness may lead to a situation where biases and unfair decisions are not being detected.

Finally, while most explanation approaches focus on explaining at the domain level, that is how much a certain feature has contributed to a certain decision, it requires more research to investigate whether explaining important elements of the underlying AI model is also important to enable explainees to take better decisions, that is how the model has derived the decision and what role the data may play in better understanding how a decision support system has come to a conclusion.

Acknowledgements The preparation of the manuscript by Britta Wrede was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): TRR 318/1 2021-438445824 and partly

by the Collaborative Research Center SFB 1320 'EASE - Everyday Activity Science and Engineering'. The prepatation of the manuscript by Ute Schmid was funded by the Deutsche Forschungsgemeinschaft (DFG) – Project number 427404493 (Dare2Del).

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Abdelghanisps R, Pierre-Yves O, Edith L, Catherine V, Hélène S (2022) Conversational agents for fostering curiosity-driven learning in children. Int J Hum Comput Stud 167:102887
2. Adadi A, Berrada M (2018) Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE Access 6:52138–52160
3. Ai L, Stephen HM, Céline H, Mark G, Ute S (2021) Beneficial and harmful explanatory machine learning. Mach Learn 110(4):695–721
4. Arjun RA, Keze W, Changsong L, Sari S-S, Hongjing L, Sinisa T, Joyce C, Song-Chun Z (2022) CX-ToM: counterfactual explanations with theory-of-mind for enhancing human trust in image recognition models. Iscience 25(1):103581
5. Ali A, Jugal KK (2018) Intelligent tutoring systems: a comprehensive historical survey with recent developments. arXiv preprint arXiv:1812.09628
6. Angluin D (1987) Learning regular sets from queries and counterexamples. Inf Comput 75(2):87–106
7. Atanasove P, Jakob GS, Christina L, Isabelle A (2020) A diagnostic study of explainability techniques for text classification. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, pp 3256–3274
8. Berg S, Kutra D, Kroeger T, Straehle CN, Kausler BX, Haubold C, Schiegg M, Ales J, Beier T, Rudy M et al (2019) Ilastik: interactive machine learning for (bio) image analysis. Nat Methods 16(12):1226–1232
9. Biundo S, Wendemuth A (2016) Companion-technology for cognitive technical systems. Künstliche Intell 30(1):71–75
10. Bruckert S, Finzel B, Schmid U (2020) The next generation of medical decision support: a roadmap toward transparent expert companions. Front Artif Intell 3:507973
11. Bruner J (1985) Child's talk: learning to use language. Child Lang Teach Therapy 1(1):111–114
12. Brusilovsky P, de Marco G, Alexander F, Pasquale L, Marco P, Giovanni S, Martijn CW (2022) Joint workshop on interfaces and human decision making for recommender systems (IntRS'22). In: Proceedings of the 16th ACM Conference on Recommender Systems, RecSys'22, New York, NY, USA. Association for Computing Machinery, pp 667–670
13. Buschmeier H, Stefan K (2014) A dynamic minimal model of the listener for feedback-based dialogue coordination. In: Proceedings of SemDial 2014, Workshop on the Semantics and Pragmatics of Dialogue
14. Bussone A, Simone S, Dympna OS (2015) The role of explanations on trust and reliance in clinical decision support systems. In: 2015 International conference on healthcare informatics, IEEE, pp 160–169
15. Chen R, Hao C, Ge H, Jie R, Quanshi Z (2019) Explaining neural networks semantically and quantitatively. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), IEEE, pp 9186–9195
16. Chi M, Siler S, Jeong H (2004) Can tutors monitor students understanding accurately? Cogn Instr 22:09
17. Chi MTH, Wylie R (2014) The ICAP framework: linking cognitive engagement to active learning outcomes. Educ Psychol 49(4):219–243
18. Clancey WJ (1983) The epistemology of a rule-based expert system-a framework for explanation. Artif Intell 20(3):215–251
19. Coleman EB, Brown AL, Rivkin ID (1997) The effect of instructional explanations on learning from scientific texts. J Learn Sci 6:347–365
20. Craven M, Jude S (1995) Extracting tree-structured representations of trained networks. Advances in neural information processing systems. Springer, Cham, p 8
21. Dai W-Z, Qiuling X, Yang Y, Zhi-Hua Z (2019) Bridging machine learning and logical reasoning by abductive learning. Advances in neural information processing systems. Springer, Cham, p 32
22. Dhurandhar A, Pin-Yu C, Ronny L, Chun-Chen T, Pai-Shun T, Karthikeyan S, Payel D (2018) Explanations based on the missing: towards contrastive explanations with pertinent negatives. In: Samy B, Hanna MW, Hugo L, Kristen G, Nicolò C-B, Roman G (eds) Advances in neural information processing systems: annual conference on neural information processing systems. NeurIPS, Montréal, pp 590–601
23. Doshi F-V, Been K (2017) Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608
24. El-Assady M, Wolfgang J, Rebecca K, Udo S, Rita S, Fabian S, Thilo S, Daniel K (2019) Towards XAI: structuring the processes of explanations. In: Proceedings of HCML Workshop at CHI'19
25. Fails JF, Dan RO Jr (2003) Interactive machine learning. In: Proceedings of the 8th International Conference on Intelligent User Interfaces, pp 39–45
26. Finzel B, David ET, Stephan S, Ute S (2021) Explanation as a process: user-centric construction of multi-level and multi-modal explanations. In: Stefan ER, Elmar R (eds) KI 2021: advances in artificial intelligence - 44th German conference on AI, virtual event, September 27 - October 1, 2021, proceedings, volume 12873 of lecture notes in computer science. Springer, Cham, pp 80–94
27. Finzel B, David ET, Anna MT, Ute S (2021) Multimodal explanations for user-centric medical decision support systems. In: HUMAN@AAAI Fall Symposium
28. Gentner D, Markman AB (1994) Structural alignment in comparison: no difference without similarity. Psychol Sci 5(3):152–158
29. Glass M (2001) Processing language input in the CIRCSIM-tutor intelligent tutoring system. Artificial intelligence in education. IOS Press, Amsterdam, pp 210–221
30. Gobert JD, Clement JJ (1999) Effects of student-generated dia- grams versus student-generated summaries on conceptual

understanding of causal and dynamic knowledge in plate tectonics. J Res Sci Teach 36:39–53
31. Goebel R, Chander A, Holzinger K, Lecue F, Akata Z, Stumpf S, Kieseberg P, Holzinger A (2018) Explainable AI: the new 42? International cross-domain conference for machine learning and knowledge extraction. Springer, Cham, pp 295–303
32. Guerdan L, Alex R, Hatice G (2021) Toward affective XAI: facial affect analysis for understanding explainable human-AI interactions. In: Proceedings of of the IEEE/CVF International Conference on Computer Vision, pp 3796–3805
33. Gunning D, Aha D (2019) DARPA's explainable artificial intelligence (XAI) program. AI Mag 40(2):44–58
34. Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang G-Z (2019) XAI - explainable artificial intelligence. Sci Robot 4(37):eaay7120
35. Hemminghaus J, Stefan K (2017) Towards adaptive social behavior generation for assistive robots using reinforcement learning. In: Bilge M, Manfred T, Astrid W, James EY (eds) Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction, HRI. ACM, Vienna, pp 332–340
36. Hempel CG, Oppenheim P (1948) Studies in the logic of explanation. Philos Sci 15:135–75
37. Herchenbach M, Müller D, Scheele S, Schmid U (2022) Explaining image classifications with near misses, near hits and prototypes. International conference on pattern recognition and artificial intelligence (ICPRAI). Springer, Cham, pp 419–430
38. Holzinger A (2016) Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Inf 3(2):119–131
39. Holzinger A (2016) Interactive machine learning (IML). Informatik-Spektrum 39(1):64–68
40. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H (2019) Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev 9(4):e1312
41. Kambhampati S (2021) Polanyi's revenge and AI's new romance with tacit knowledge. Commun ACM 64(2):31–32
42. Keil FC (2006) Explanation and understanding. Annu Rev Psychol 57:227–254
43. Kim B, Khanna R, Koyejo OO (2016) Examples are not enough, learn to criticize! Criticism for interpretability. Advances in neural information processing systems. Springer, Cham, p 29
44. Keiichi K (2020) Learning by teaching face-to-face: the contributions of preparing-to-teach, initial-explanation, and interaction phases. Eur J Psychol Educ 37:551–566
45. Kozok B (1998) Erklärungsfunktionalität von expertensystemen: Anwendungs- und forschungsbedarf. Künstliche Intell 12(1):25–33
46. Kuehl N, Jodie L, Christian M (2020) Do you comply with AI? Personalized explanations of learning algorithms and their impact on employees' compliance behavior. arXiv:2002.0877
47. Kulesza T, Margaret B, Weng-Keen W, Simone S (2015) Principles of explanatory debugging to personalize interactive machine learning. In: Proceedings of 20th International Conference on Intelligent User Interfaces, pp 126–137
48. Laato S, Tiainen M, Najmul Islam AKM, Mäntymäki M (2021) How to explain AI systems to end users: a systematic literature review and research agenda. Internet Res 32(7):1–31
49. Lakkaraju H, Stephen HB, Jure L (2016) Interpretable decision sets: a joint framework for description and prediction. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 1675–1684
50. Langer M, Oster D, Speith T, Hermanns H, Kästner L, Schmidt E, Sesing A, Baum K (2021) What do we want from Explainable Artificial Intelligence (XAI)? A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artif Intell 296:103473
51. Langley C, Bogdan-Ionut C, Fabio C, Barbara JS (2022) Editorial: theory of mind in humans and in machines. Front Artif Intell 5:917565
52. Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller K-R (2019) Unmasking Clever Hans predictors and assessing what machines really learn. Nat Commun 10(1):1–8
53. Lombrozo T (2007) Simplicity and probability in causal explanation. Cogn Psychol 55(3):232–257
54. Michie D (1988) Machine learning in the next five years. Proceedings of the third European working session on learning. Pitman, Cham, pp 107–122
55. Martijn M, Toon W, Katrien V (2021) Your eyes explain everything: exploring the use of eye tracking to provide explanations on-the-fly. In: Proceedings of 8th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems co-located with 15th ACM Conference on Recommender Systems (RecSys 2021), volume 2948
56. Miller T (2019) Explanation in artificial intelligence: Insights from the social sciences. Artif Intell 267:1–38
57. Mohseni S, Zarei N, Ragan ED (2021) A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ACM Trans Interact Intell Syst 11(3–4):1–45
58. Molnar C (2020) Interpretable machine learning. Lulu.com
59. Muggleton SH, Schmid U, Zeller C, Tamaddoni-Nezhad A, Besold T (2018) Ultra-strong machine learning: comprehensibility of programs learned with ILP. Mach Learn 107(7):1119–1140
60. Mueller D, Michael M, Stephan S, Ute S (2022) An interactive explanatory AI system for industrial quality control. Thirty-sixth AAAI conference on artificial intelligence, AAAI 2022, thirty-fourth conference on innovative applications of artificial intelligence, IAAI 2022, the twelveth symposium on educational advances in artificial intelligence, EAAI 2022 virtual event. AAAI Press, London, pp 12580–12586
61. Nguyen A, Yosinski J, Clune J (2019) Understanding neural networks via feature visualization: a survey. In: Samek W et al (eds) Explainable AI, LNAI 11700. Springer, Cham, pp 55–76
62. Nwana HS (1990) Intelligent tutoring systems: an overview. Artif Intell Rev 4(4):251–277
63. Quasthoff U, Heller V, Morek M (2017) On the sequential organization and genre-orientation of discourse units in interaction: an analytic framework. Discourse Stud 19(1):84–110
64. Rabold J, Gesina S, Ute S (2020) Expressive explanations of DNNs by combining concept analysis with ILP. In: Ute S, Franziska K, Diedrich W (eds) KI 2020: Advances in artificial intelligence – 43rd German conference on AI, Bamberg, Germany, september 21-25, 2020, proceedings, volume 12325 of lecture notes in computer science. Springer, Cham, pp 148–162
65. Rabold J, Siebers M, Schmid U (2018) Explaining black-box classifiers with ILP-empowering LIME with Aleph to approximate non-linear decisions with relational rules. International conference on inductive logic programming. Springer, Cham, pp 105–117
66. Rabold J, Siebers M, Schmid U (2022) Generating contrastive explanations for inductive logic programming based on a near miss approach. Mach Learn 111(5):1799–1820
67. Ribeiro TM, Sameer S, Carlos G (2016) 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, pp 1135–1144
68. Rips LJ (2002) Circular reasoning. Cogn Sci 26:767–795
69. Rohlfing KJ, Cimiano P, Scharlau I, Matzner T, Buhl HM, Buschmeier H, Esposito E, Grimminger A, Hammer B, Häb-Umbach R, Horwath I, Hüllermeier E, Kern F, Kopp S, Thommes K, Ngomo A-CN, 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 Trans Cogn Dev Syst 13(3):717–728

70. Roth-Berghofer T, Richter MM (2008) Gasteditorial erklärungen. Künstliche Intell 22(2):4
71. Rozenblit LR, Keil FC (2002) The misunderstood limits of folk science: an illusion of explanatory depth. Cogn Sci 26:521–562
72. Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215
73. Stuart R (2019) Human compatible: artificial intelligence and the problem of control. Penguin, London
74. Samek W, Montavon G, Vedaldi A, Hansen LK, Müller K-R (2019) Explainable AI: interpreting, explaining and visualizing deep learning, vol 11700. Springer Nature, Cham
75. Schallner L, Rabold J, Scholz O, Schmid U (2019) Effect of superpixel aggregation on explanations in LIME: a case study with biological data. In: Cellier P, Driessens K (eds) Machine learning and knowledge discovery in databases – international workshops of ECML PKDD 2019, Würzburg, Germany, september 16-20, 2019, proceedings, part i, volume 1167 of communications in computer and information science. Springer, Cham, pp 147–158
76. Schmid U (2018) Inductive programming as approach to comprehensible machine learning. In: Beierle C, Kern-Isberner G, Ragni M, Stolzenburg F, Thimm M (eds) Proceedings of 7th workshop on dynamics of knowledge and belief (DKB-2018) and the 6th workshop KI and kognition (KIK-2018) co-located with 41st German conference on artificial intelligence (KI 2018), Berlin, Germany, September 25, 2018, volume 2194 of CEUR workshop proceedings. CEUR-WS.org, Berlin, pp 4–12
77. Schmid U (2021) Interactive learning with mutual explanations in relational domains. In: Muggleton S, Charter N (eds) Human-like machine intelligence. Oxford University Press, Oxford, pp 337–353
78. Schmid U, Finzel B (2020) Mutual explanations for cooperative decision making in medicine. KI-Künstliche Intell 34(2):227–233
79. Schmid U, Tresp V, Bethge M, Kersting K, Stiefelhagen R (2020) Künstliche intelligenz - die dritte welle. In: Reussner RH, Koziolek A, Heinrich R (eds) Jahrestagung der Gesellschaft für Informatik, INFORMATIK 2020 - back to the future, Karlsruhe, Germany, 28. september - 2. Oktober 2020, volume P-307 of LNI. GI, Karlsruhe, pp 91–95
80. Siebers M, Schmid U (2019) Please delete that! Why should I? Explaining learned irrelevance classifications of digital objects. KI-Künstliche Intell 33(1):35–44
81. Sokol K, Flach P (2020) Explainability fact sheets: a framework for systematic assessment of explainable approaches. In: Proceedings of 2020 Conference on Fairness, Accountability, and Transparency, pp 56—67
82. Teso S, Hinz O (2020) Challenges in interactive machine learning. KI-Künstliche Intell 34(2):127–130
83. Teso S, Kristian K (2019) Explanatory interactive machine learning. In: Proceedings of 2019 AAAI/ACM Conf on AI, Ethics, and Society, pp 239–245
84. Thaler A, Ute S (2022) Enriching intelligent tutoring systems (ITS) with methods of explainable artificial intelligence (XA) – a research proposal. Reflection on intelligent systems: towards a cross-disciplinary definition, Stuttgart
85. Thaler AM, Ute S (2021) Explaining machine learned relational concepts in visual domains-effects of perceived accuracy on joint performance and trust. Proc Annu Meet Cogn Sci Soc 43:1705–1711
86. Tintarev N, Judith M (2007) A survey of explanations in recommender systems. In: 2007 IEEE 23rd International Conference on Data Engineering Workshop, IEEE, pp 801–810
87. Tjoa E, Guan C (2021) A survey on explainable artificial intelligence (XAI): toward medical XAI. IEEE Trans Neural Netw Learn Syst 32(11):4793–4813
88. VanLehn K (2011) The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educ Psychol 46(4):197–221
89. Wachter S, Mittelstadt B, Russell C (2017) Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harv JL Tech 31:841
90. Wang D, Yang Q, Abdul A, Lim BY (2019) Designing theory-driven user-centric explainable AI. Association for Computing Machinery, New York, pp 1–15
91. Wang X, Shengcheng Y, Hui Z, Michael L, Katia S (2019) Verbal explanations for deep reinforcement learning neural networks with attention on extracted features. In: 2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), IEEE, pp 1–7
92. Weitz K (2021) Towards human-centered AI: psychological concepts as foundation for empirical XAI research. it - Information Technology
93. Wood D, Bruner JS, Ross G (1976) The role of tutoring in problem solving. J Child Psychol Psychiatry 17(2):89–100
94. Yeh C-K, Hsieh C-Y, Suggala AS, Inouye DI, Ravikumar P (2019) On the (in)fidelity and sensitivity of explanations. In: Wallach HM, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox EB, Garnett R (eds) Advances in neural information processing systems 32: annual conference on neural information processing systems. NeurIPS, Vancouver, pp 10965–10976
95. Zeller C, Ute S (2016) Automatic generation of analogous problems to help resolving misconceptions in an intelligent tutor system for written subtraction. In: Coman A, Kapetanakis S (eds) Workshops proceedings for the 24th international conference on case-based reasoning (ICCBR 2016), vol 1815. CEUR workshop proceedings. CEUR-WS.org, Berlin, pp 108–117
96. Zhou J, Gandomi AH, Chen F, Holzinger A (2021) Evaluating the quality of machine learning explanations: a survey on methods and metrics. Electronics 10(5):593