Requirements for Explainability and Acceptance of Artificial Intelligence in Collaborative Work

Sabine Theis[1][0000−0002−3422−3734], Sophie Jentzsch[1][0000−0001−6217−8814], Fotini Deligiannaki[2][0000−0002−9479−8767], Charles Berro[2][0000−0002−2662−8774], Arne Peter Raulf[2][0009−0003−8672−3014], and Carmen Bruder[3][0000−0003−2638−2361]

[1] Institute for Software Technology, Linder Höhe, 51147 Cologne
[2] Institute for AI Safety and Security, Rathausallee 12, 53757 Sankt Augustin
[3] Institute for Aerospace Medicine, Sportallee 5a, 22335 Hamburg
{firstname}.{lastname}@DLR.de

Abstract. The increasing prevalence of Artificial Intelligence (AI) in safety-critical contexts such as air-traffic control leads to systems that are practical and efficient, and to some extent explainable to humans to be trusted and accepted. The present structured literature analysis examines $n = 236$ articles on the requirements for the explainability and acceptance of AI. Results include a comprehensive review of $n = 48$ articles on information people need to perceive an AI as explainable, the information needed to accept an AI, and representation and interaction methods promoting trust in an AI. Results indicate that the two main groups of users are developers who require information about the internal operations of the model and end users who require information about AI results or behavior. Users' information needs vary in specificity, complexity, and urgency and must consider context, domain knowledge, and the user's cognitive resources. The acceptance of AI systems depends on information about the system's functions and performance, privacy and ethical considerations, as well as goal-supporting information tailored to individual preferences and information to establish trust in the system. Information about the system's limitations and potential failures can increase acceptance and trust. Trusted interaction methods are human-like, including natural language, speech, text, and visual representations such as graphs, charts, and animations. Our results have significant implications for future human-centric AI systems being developed. Thus, they are suitable as input for further application-specific investigations of user needs.

Keywords: Artificial intelligence · Explainability · Acceptance · Safety-critical contexts · Air-traffic control · Structured literature analysis · Information needs · User requirement analysis

1 Introduction

Humans will collaborate with *artificial intelligence (AI)* systems in future living and working environments. In particular, this will characterize aviation, medicine or space travel activities. These outstanding safety-critical application areas require – more than others – the consideration of individual requirements of human operators in the design of collaborative assistance systems. In this context, the German Aerospace Center (DLR) is developing guidelines for the human-centered collaboration design between users and AI systems. The focus is on tasks and contexts where operators, such as *air traffic controllers (ATCOs)*, medical professionals, or operators of space systems work collaboratively with AI to achieve efficient and safe operation. Especially humans as the operators of AI systems form a thematic priority, together with the question of how explainability and acceptance can be assured for the users of AI systems. To develop an explainable and acceptable AI pilot and AI air traffic co-controller, this article examines previously noted *user requirements* in collaboration with artificial intelligence. In general, *user requirement analysis* denotes an iterative process in which one identifies [80], specifies, and validates functional and non-functional characteristics of an IT system [40, 66] together with individual users and user groups. The main goal of this user requirement engineering process is to ensure that the system to be developed meets the needs of its users [67, 10]. This requires a deep understanding of user characteristics, goals, and tasks, as well as the context in which the software or system will be used [26, 71]. User requirement engineering is a critical part of the software development process, as it can significantly impact the success or failure of the final product [40].

One variable within the context of requirements engineering of data-driven systems refers to the data and information that a user requires in order to achieve their goals or perform their tasks [103, 102, 99]. A focus on *users' information needs (IN)* and seeking behavior during user requirement engineering focuses the development process less on technology and more on the essential result of information technology, namely the transfer of information to the human, which is especially relevant for human-AI interaction or data visualization systems [18, 100]. Information needs essentially describe the gap between a current and the desired state of knowledge that needs to be filled to achieve a goal which in the present case is to understand an AI [114]. IN is defined by an individual and can vary in specificity, complexity, and urgency while being induced by social, affective, and cognitive needs.

The main contribution of the present article is a brief and understandable overview of the technical background of AI, *explainable artificial intelligence (XAI)*, and *human-in-the-loop (HITL)* in AI development. Through an interdisciplinary synthesis of computer science and psychological knowledge the present article address the need for human-centered explainable and trusted artificial intelligence by eliciting user requirements through a structured analysis of previous work on human-AI interaction.

2 Background

The following paragraphs briefly introduce related technological aspects to establish a common understanding of the broader context. First, challenges and state-of-the-art of XAI are discussed from a technical perspective in Sec. 2.1. Sec. 2.2 then provides a human perspective on explainability in intelligent systems. Finally, Sec. 2.3 brings both perspectives together by discussing the HITL paradigm.

2.1 Technical Perspective

AI is an umbrella term for a wide variety of different systems and techniques. When AI first emerged in the 1950s, the main focus was on symbolic AI, where real-world concepts are represented by symbolic entities, and human behavior is expressed by explicitly formulated logical rules. However, with exponentially growing computational resources and a stronger focus on statistical approaches, AI went through a sharp paradigm shift in the 1990's: from a logic-based to a data- and representation-driven doctrine [87]. This was the start of the era of *machine learning (ML)* and later *deep learning (DL)*, which is a subdomain of ML using (deep) neural networks [73]. ML has reached multiple milestones in various domains [85, 86, 7], exploiting massive amounts of data with sophisticated algorithms.

Nowadays, the field of AI is strongly dominated by ML and DL, turning the vast majority of concrete XAI implementations towards ML-based sub-problems. Understanding ML-based systems is especially challenging and relevant for several reasons, outlined in the following paragraphs.

ML and explainability XAI is a constantly growing research field, yet there exists no single established definition of explainability and related concepts such as transparency or interpretability [6]. The EASA defines *explainability* as the "*capability to provide the human with understandable and relevant information on how an AI/ML application is coming to its results.*" [36]. A strongly related term in literature is *interpretability* as it is often defined similarly, e.g., the users' ability to "*correctly and efficiently predict the method's results*" [60].

Achieving system explainability is a relevant requirement for numerous reasons, essentially though because ML-based systems' decisions affect many aspects of our daily lives and need to be proven reliable. While empirical evidence on the effects of system explainability on users' trust remains inconclusive, explainability certainly supports *trustworthiness* [58]. In addition, Gerlings et al. outline a comprehensive list of motivations for XAI which are, among others, generation of trust and transparency, following compliance and regulations (e.g. GDPR), social responsibility and risk avoidance [44].

Although it is evident that XAI is a fundamental requirement it comes along with multiple technical and systemic obstacles. In a standard ML pipeline, enormous amounts of data are fed into an algorithm that autonomously identifies and encodes relevant patterns into a model. In contrast to symbolic AI, state-of-the-art ML models encode real-world information and human knowledge implicitly into inherently opaque models. It is, therefore, especially challenging to retrace their latent reasoning and portraying their insights rationally, which is why they are frequently referred to as *black boxes*.

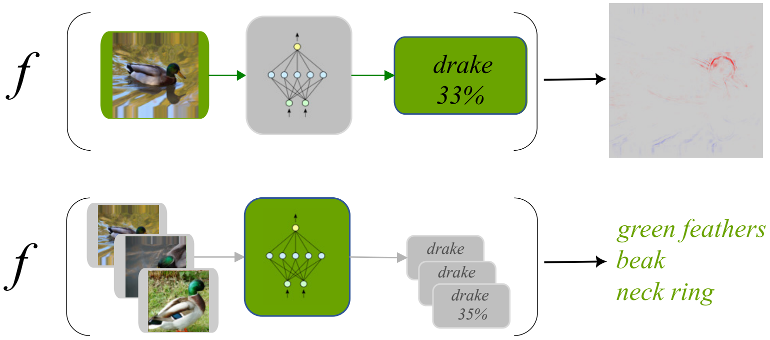

Fig. 1. The concept of a local (top) and global (bottom) explanation function f is depicted. LRP [77], the former explanation method[2], solely depends on the specific input/output pair. The output highlights the useful to the network image area (as seen in red, the duck's head). The later yields abstract features/concepts, targets the AI model in itself and explains what features are used in its decision. The input images are taken from the ImageNet data set [31].

Explanation characteristics Technically, explanation approaches can be divided by a row of characteristics [36, 75]:

Global vs. local While global explanation approaches seek to describe the overall model, answering the question *what information does the model utilize to answer?*, local approaches are designed to explain specific model outputs or the role of particular input samples, i.e, *why did the model yield a specific output from a specific input?*. Both approaches are illustrated in Fig. 1.

Model-specific vs. model-agnostic Depending on the specific system or the types of input data, different explanation techniques can be appropriate. Other approaches claim to be model agnostic. That means they can be applied to every type of underlying ML model.

Intrinsic vs. post-hoc explainability Some ML models, e.g., linear and logistic regression or decision trees, are intrinsically interpretable. These models can be explained by restricting the models' complexity [76]. In contrast, models that do not possess that characteristic can be explained through so-called post-hoc approaches, i.e, generating explanations for contemplation after the training process [56].

Explainee The explainee, i.e. the recipient of the explanation, plays a central role; consequently, system's explanations require a different level of detail for an expert or a developer compared to a naive, non-expert user.

[2] The following open API was used for generating explanations: https://lrpserver.hhi.fraunhofer.de/image-classification

Tools and approaches for XAI In the ML community, a row of explanation approaches has been established recently. These approaches mainly bear on ML models not intrinsically interpretable [76], such as deep neural networks.
Many XAI methods aim at highlighting the most relevant (to a certain outcome) features of the input data. In the case of neural networks, *Layer-wise relevance propagation (LRP)* [77] works by propagating the prediction backward through the system and can be used to unmask correct predictions being made for the wrong reasons [68]. Similarly, *LIME* and *SHAP* are python data visualization libraries. All mentioned approaches generate local post-hoc explanations and are model-agnostic.
A method partially related to unveiling correct decision being taken for false reasons are *counterfactual explanations*, that determine and highlight which features need to be different to receive a different system outcome [110]. *Concept-based* explanations aim to identify relevant higher-level concepts instead of features specific to the input data. As such, they focus on meaningful human concepts, establishing human-understandable explanations [46, 61]. A non-technical measure to enhance the explainability and responsible deployment of intelligent systems is the convention of *model cards*. This framework specifies relevant details regarding the model's training, evaluation, and intended usage, which helps practitioners to understand the context and conclude assumptions about inner workings [72]. The fusion of modern ML approaches with symbolic AI yields methods depicting learned representations from neural networks symbolically in an inherently intuitive structure.They appear highly effective for achieving interpretability, trust and reasoning (also see Sec. 5). Primary methods of *neural-symbolic learning* [117] aim at injecting semantics, as seen in [1], or expert knowledge in the form of knowledge graphs [34].

2.2 Human Perspective

Following the preceding description of the technical perspective on the collaboration between humans and artificial intelligence, this section provides an overview of important concepts from human factors research on the collaboration and coordination between human operators and AI in domains where safety is critical.

Collaboration at work *Collaboration* is based on the human's ability to participate with others in collaborative activities with shared goals and intentions, as well as the human's need to share emotions, experiences, and activities with each other [104]. This enables people to work together and understand each other. As a consequence, human-centered integration of AI should address humans' expectations on their human partners as well as digital partners.
In domains where safety is of critical importance and human error can have severe consequences [51, 89], human operators often work together in control centers [98] to achieve efficient and safe operation. Examples are airport operational centers, air traffic control centers, nuclear power plants, and military control centers. In control centers, teams of human operators have to work under time pressure to

supervise complex dynamic processes as well as decide for remedy. Supervisory control is the human activity involved in initiating, monitoring, and adjusting processes in systems that are otherwise automatically controlled [24]. Being a supervisor takes the operator out of the inner control loop for short periods or even for significantly longer periods, depending on the level at which the supervisor chooses to operate [112]. Workshops with experienced pilots and ATCOs, which were conducted in order to gather their expectations about future tasks, roles, and responsibilities, indicated that task allocation, teamwork, and monitoring in a highly automated workplace pose challenges [16]. As supervisory control is one of the core tasks in control rooms, teams of operators are required to monitor the systems appropriately [91]. Through interactions, operators in a team can dynamically modify each other's perceptual and active capabilities [49]. However, when monitoring a system, it is essential that human operators work together effectively and cooperatively [23, 89]. With this in mind, communication in control operations is of high importance. Communication as a "meta-teamwork process that enables the other processes" [82] provides indications for the coordinative activities while monitoring.Especially in critical situations, "it is not only critical that teams correctly assess the state of the environment and take action, but how this is accomplished" [22]. As a consequence, a team's communication provides insight into how the team members deal with critical situations.

Trust and acceptance Especially in domains where safety is of critical importance, both *trust* of the users in human-human interactions and trust in human-technology interactions is of vital importance [14]. Trust as a psychological concept is defined as a belief in the reliability, truth, ability, or strength of someone or something [5]. Trust influences interpersonal relationships and interaction and plays a fundamental role in decision-making [33] and risk perception [35]. Trust can be influenced by past experiences [20], communication [13], and behaviors.

Trust in automation can be conceptualized as a three-factor model consisting of the human trustee, the automated trustee, and the environment or context. In this model, qualities of the human (such as experience), work with qualities of the autonomous agent (such as form) in an environment that also influences the nature of the interaction. Since trust is constantly evolving, time itself is also a facet of trust in human-automation interactions. Measurement of trust is challenging because trust itself is a latent variable, and not directly observable. [65]. To make the complexity of the concept more manageable, technical perspectives often consider it as the extent to which a human believes the AI's outputs are correct and useful for achieving their current goals in the current situation [105]. Trust and *acceptance* are related in that a person is more likely to accept something if they trust it, which is investigated for users' trust in AI technologies by [21]. Trust can provide a sense of confidence and security, which can make it easier for a person to accept something. In addition to the concept of trust, human acceptance of technology plays an important role. Technology acceptance is the extent to which individuals are willing to use and adopt new technological inno-

vations [93, 28]. It is a multi-dimensional concept that takes into account various factors that influence an individual's decision to use a particular technology. The concept of technology acceptance is rooted in the theory of reasoned action [50] and the theory of planned behavior [3]. Technology acceptance is influenced by a range of factors, including the perceived usefulness and perceived ease of use of the technology, social influence, trust, compatibility with existing technologies and practices, perceived risks, and anxiety about using the technology.
To summarize the human perspective, it can be stated that a successful integration of AI in control centers' operations has to consider humans' expectations on their human partners and digital partners. To address the humans' expectations on their human partners and digital partners, AI systems should:(1) support teamwork in in safety critical situations, (2) facilitate situation awareness, (3) consider the requirements of supervisory control, (4) support communication between team members, and (5) consider both interpersonal trust and trust in technology.

2.3 Human-in-the-loop Methods

The information flow in AI-based automated systems can be represented as a loop: the *environment* is recorded using sensors; the *data* produced by the recordings is consumed by the *algorithm* to either train a *model* or to use the model to infer a *result*; the result is used as a command for an automation to modify the environment. In order to trust the system in safety critical applications, humans must have the oversight and understanding of the various elements of the loop. It is therefore natural to place the human in the loop (also see Fig. 2).

Definition The *Human-in-the-loop* (HITL) paradigm is a set of human oversight mechanisms on systems running AI models. Such mechanisms implement human-computer interaction methods at different levels of the AI-based system life-cycle such as data collection, model design, training process, model evaluation or model inference [39, 116, 25]. Overall, HITL brings together research fields from computer science, cognitive science and psychology [116].

Approaches The implementation of HITL in a ML system primarily depends on the degree and nature of human knowledge to be injected. This can take place throughout the entire ML pipeline as seen in Fig. 2. Following the categorization in [116] an initial approach is performing data processing with HITL. The goal is obtaining a valid data set, i.e. which is accurately labeled (with the help of human annotators) essentially at key/representative data samples, stemming from a pool of unlabeled data. The above method employs expert feedback before the ML model training and inference take place. During training, feedback can be used to push the model to map its knowledge as closely as possible to humans' (i.e, rewarding alignment with their decisions) or by learning through imitation [25] in the case of *reinforcement learning (RL)* agents. Finally, HITL coupled with model inference is best described by the application areas where

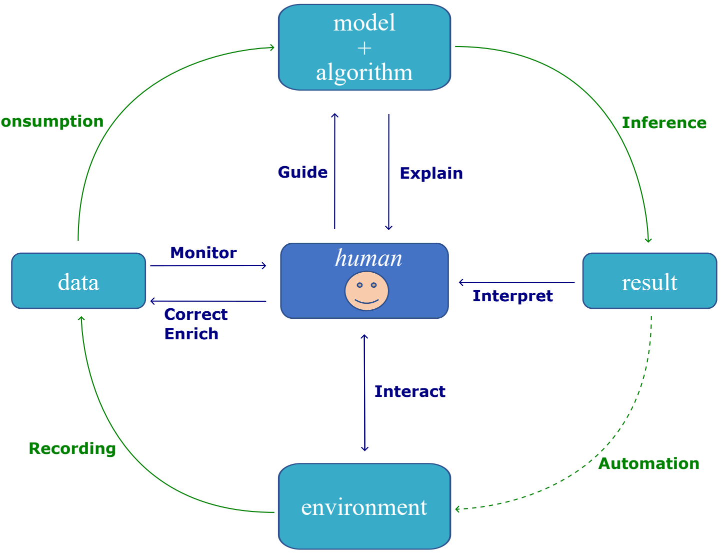

Fig. 2. A figure depicting the data/information/knowledge/command flow – showed as directional arrows – in a classical *human-in-the-loop* approach. The AI-based system is abstractly depicted as consisting of the data – recorded from the environment – given as training/inference input; the model architecture and learning algorithm; followed lastly by its results (dependent on the AI's task). The results might be used as commands for an automated system to act on the environment. The blue arrows should be interpreted as the human - either a developer or a non-expert user - performing the action (e.g. "the human guides the model", or "the human interprets the result").

the outcomes of ML models are used or processed by humans (occasionally interactively and iteratively). Collaboration is mostly imagined in this setting since the human has multiple abilities to interact with the outputs such as choosing between or observing multiple outcomes, to ask for explanations on what they represent or to further refine them by accepting/rejecting the AI's assistive input [7].

Applications A HITL system shows great results in domains where the human creativity and fine understanding of the context is combined with the machine's data-analysis to reach performance superior to the human alone or the machine alone [7]. Trust and acceptance can be built as well, as seen in [106] where HITL for data labeling is employed to improve automatic conflict resolution suggestions within the air traffic management. Specifically, claiming that modern methods are not fully trusted by human operators, such as ATCOs and pilots, the authors enhance their acceptance by combining human-generated resolutions with RL algorithms.
The collaboration of human and AI-based systems – as a hybrid team – is particularly relevant in safety critical applications where the strengths of the machine

on data-analysis and tasks repetitiveness are combined to the context understanding and adaptability to new scenarios of the human operator. As in a *4-Eyes Principle* team organization, it is expected for the hybrid team to be less prone to missing relevant information or to overlooking effective solutions. On top of that, in collaborative RL schemes the safety of the human teaching the AI-based system is naturally prioritized. Consequently, random and/or dangerous actions of this system can be mitigated by sophisticated on-the-fly guidance from humans, as noted in [37].
Concerning the current and future focus in HITL systems, the authors in [116] indicate that existing methods need to learn more effectively from expert human experience, essentially by moving towards more complex and less simplistic and superficial human intervention.

3 Problem and Research Questions

XAI techniques provide information that describes how ML or DL models generate results based on data or processes. Depending on the model type, its results can be explained either by reducing complexity or by looking back at its training or evaluation. These techniques aim to demonstrate the effectiveness of models for developers of ML and DL models. However, to what extent information resulting from XAI methods aligns with the users' requirements remains unclear. Although HITL approaches allow for human input, their purpose rather addresses improving the life cycle of the AI model, including its data collection, model design, training, and evaluation. So far, less attention is given to explanations serving the user and contextual goals. Human-computer interaction and user-centered design have long addressed the challenges of developing technical systems that meet user needs. Eliciting the requirements of different user groups may provide valuable insights for developing accepted and trusted AI. In the run-up to requirements analysis, the following research question arises:

RQ1: What information do people need to perceive an AI system as explainable or understandable?

RQ2: What information do people need to accept an AI system?

RQ3: Which interaction/ information representation methods are trustworthy?

4 Method

In order to answer the aforementioned research questions, a structured literature review was conducted. Its methodological procedure is described in the following section, following the general methodological framework of the *Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA)* statement published in 2009 [74] and updated in 2020 [81].

4.1 Databases and Search Query

Accordingly, a comprehensive literature search was conducted in the "Web of Science" and "Google Scholar" databases, the DLR repository "eLib," the "DLR Library Catalog", the "NASA Technical Report Server", as well as the "Ebook Central" portal, and the database of German national libraries. We identified search terms and used Boolean operators to generate the following query strings for searching each of the mentioned sources the search term combination (("explainability" OR "traceability" OR "acceptance") AND ("artificial intelligence") AND ("reasoning" OR "problem solving" OR "knowledge representation" OR "automatic planning" OR "automatic scheduling" OR "machine perception" OR "computer vision" OR "robotics" OR "affective computing")).

4.2 Identification and Screening

Here, relevant articles in mentioned data sources were identified from 01.08.22 to 08.10.22. In total, $n = 244$ articles identified as relevant were returned from the "Web of Science Core Collection", $n = 240$ relevant articles from a total of $n = 18,700$ Google scholar results, and $n = 27$ from the DLR search consisting of $n = 16$ from NASA Technical Reports, $n = 5$ from eLib publications, and $n = 6$ articles originating from the DLR library catalog. The latter were not included because source details were not available. The search results of all three queries were saved to .ris files and imported into the browser-based literature management program Paperpile, and data source tags were added here. When exporting the Web of Science Core Collection results, there was a loss of $n = 10$ records that were presumed duplicates. In addition, another record was recognized as a duplicate in Paperpile and removed from the initial records set. This resulted in $n = 521$ initially identified records as input for screening.

4.3 Inclusion and Exclusion Criteria

During the screening, each reference was screened first by machine filtering and then by manually checking the titles and abstracts for the following a priori inclusion and exclusion criteria: (1) the language of the article is in English, (2) the publication date of the article is between the years 1950 and 2022 inclusive, (3) the article contains results on explainability, and user acceptance of systems where humans interact and collaborate with AI, (4) the AI systems perform at least one of the following tasks: reasoning, problem-solving, knowledge representation, automated planning, and scheduling, ML, natural language processing , machine perception (computer vision), machine motion and manipulation (robotics), or emotional or social intelligence (affective computing). After excluding $n = 402$, banned records $n = 109$ could be retrieved and assessed in detail for mentioned inclusion and exclusion criteria.

4.4 Content Assessment

In the subsequent systematic review of $n = 48$ reports, two different content assessments were made: (a) the implicit or explicit perspective of humans and (b) the quality of evidence for the outcome of interest. In addition, qualitative narrative analysis and synthesis through tabulation were performed. Of the 48 articles subjected to close examination, 32 were journals, 13 were conference papers, and 3 were technical reports—articles from 1980–2022. While early articles focus on describing technical implementation and function, the proportion of empirical evaluation of technical systems and inclusion of the user perspective increases over time.

5 Results

The following section presents the key findings of previously described literature review, highlighting the information needed for explainability and acceptance together with trustworthy methods for humans interaction with AI systems, as well as trustworthy information representation methods.

5.1 Information Needed for Explainability

In this section, we explore the information needs of different users in understanding the results and behavior of AI systems. Analysis additionally revealed characteristics of explanations that are most effective in supporting human reasoning.

Information explaining the model A much-regarded user group in the development of explainable AI are its developers, who have extensive technical and AI-specific background knowledge [41, 84, 11, 12, 90, 30]. For AI development, developers require information to understand the data and the model in terms of its internal operations, such as the weighting of individual parameters, features, or nodes, as well as information about the relation between input and output variables [27, 56]. Such model-specific information can largely be generated through model-specific, global, and local XAI methods, as described in the background section of this paper. In addition, the contextual information of the use case or the development process may be important. However, these are rarely addressed by common XAI methods [27, 56, 111] and require at least HITL approaches. While model-specific XAI can enhance the explainability of AI techniques to developers [97, 47, 43], these effects do not necessarily carry over to non-expert end users we subsequently shed light onto their requirements. However, explaining an AI through accessible raw data, code, or details about the AI models can also entail disadvantages, such as code manipulation and restrictions on the inventor's potential for innovation. In this regard, a balance needs to be found between the need for transparency and the demand for ownership [107]. [107] argue in their position paper to put more effort into understanding the requirements of all relevant user groups of an AI system to ensure that information for explaining the AI can be understood and thereby increase efficiency and effectiveness of the system.

Information explaining results Most important information non-expert users require is information explaining the results as output or behavior of an AI system [63, 32, 94, 119, 53] by answering the “WHY?”-question. This is often addressed by providing the raw data from which the results were derived from [111, 94, 8, 79, 92, 15] as well as through access to additional data used to generate the results, such as user interaction data [119]. Beyond that, users require information that explains why specific properties are assigned to certain result [27, 4, 53, 57] and to which extent features [115, 27], rules and decisions contributed to a specific result [19]. As with technical users, non-technical users need model-specific statistical information, data sources, and their biases or quality in terms of six dimensions, which are completeness, uniqueness, timeliness, validity, accuracy, and consistency [57, 17], and algorithmic information [53].

Characteristics of understandable explanations Information that explains the results or actions of an AI is particularly effective for humans if it supports logical inductive and deductive reasoning [119, 4, 111, 53]. Humans generally understand explanatory information best if they are presented in a contrasting manner [108, 57, 17]. Thereby, different properties of a result, different results with other properties, and results with different properties at different points in time should be compared to each other [27].

However, most model-agnostic, local XAI methods provide far more detail than what an end-user requires for a satisfactory explanation [19]. Explanatory information should therefore include various levels of detail represented conditionally to context, explanation capability of the AI, and temporal, perceptual, and cognitive resources of the user.

Contextual information, domain knowledge, and meta-information such as time and location are vital for domain experts, e.g., in healthcare [42]. Regarding security-critical scenarios, information should be communicated at conflict-free and less work-intensive times [63]. When representing model-specific statistics often discrete in nature, it has to be considered that human understanding and explanation of phenomena invariably utilize categories together with relationships among them, describing, for example, the relation between model predictors and the target as the relationship between entities and the target [70]. Such relationships might contain probabilistic information even if these are not as crucial for humans as causal links [108, 17]. Recent work even points out that most humans struggle to deal with uncertain information [17]. Causal information, in turn, supports humans, especially in decision-making in unfamiliar situations, but it has to be considered when individuals have prior experience with a domain, causal information can reduce confidence and lead to less accurate decisions [120, 17].

As humans prefer rare events, explanations should focus on odd reasons and be concise, meaning that shorter explanations are not considered interpretive. Form and explanation content interact largely with what is understandable [17]. To make it even more challenging, relevant contextual information extends to the person's social context, considering assumptions about the users' beliefs about themselves and their environment [17]. When including contextual information in an explanation, different users and situations have to be considered [59].

Application-specific information needs Since 2019, consideration of the user perspective has been increasing in the development of XAI. The outcome of a conversational agent supporting criminal investigations, for example, revealed that investigators want to have a clear understanding of the data, system processes, and constraints to make informed decisions and continue the investigation effectively [52].
Furthermore, different user perspectives of autonomous surface vehicles (ASV) included AI developers, engineers, and expert users, who required information about the ASV models and training data used. Operators, crew, and safety attendants wanted to get information about the current state and intention of the ASV, as well as the definition of the AI-human control boundary and when to intervene. Passengers instead needed confirmation that the ASV can see and avoid collisions with other objects [109].
Last but not least, domain experts' and lay users' trust in a robotic AI system increased by providing relevant reasons for each of its decision together with explanations of the systems' autonomous policy and the underlying reinforcement learning model through natural language question-answer dialogue [55].

5.2 Information Needed for Acceptance

To increase the chances of humans accepting an AI, it is essential to understand what information they require for acceptance. The acceptance of artificial intelligence (AI) systems is considered a proxy measure for trust [111], but can also emerge as a barrier to it [109]. Further details about the relation of both concepts are described in the Background section.

Goal-supporting information Literature suggests that information for acceptance strongly relates to the system's functions and performance, demonstrating and emphasizing the use so that the perceived usefulness of an AI system increases [63, 54]. Information supporting usefulness of a system is the goal-supporting information needed to successfully complete tasks contributing to the user's goal. For example, if an AI as part of a guidance and control system for a spacecraft provides erroneous information about system states, key functions cannot be performed, resulting in direct user rejection of the system [64]. In case of an AI recommending health decisions, correct general medical and patient-specific information together with best practice procedures are required [94] while for the acceptance of air traffic control systems, goal-supporting information include route information, air traffic information, sequential position and velocity information of other vehicles, clearance, events, vehicle responses, altitude, position of own vehicle, positions of other aircraft or information for contingencies, such as diagnosis of vehicle subsystems [69]. As with information for explainability, goal-supporting information strongly depends on the domain, task, and context of an AI system [96, 94]. However, it has been shown that across application domains, the acceptance of an AI system can be increased by making the scope and limitations of AI methods and information about potential system failures [64] known to users beforehand [29, 96].

Reliability information The acceptance of AI results benefits from attached quality or reliability indicators such as error margins, uncertainties or confidence intervals, especially in high-stakes contexts [111, 96]. All information provided must be tailored to their individual preferences and should, in general, include how data about the user is collected and processed, and how privacy is protected [119, 111]. In particular, information about the extent to which other users trust the system plays a vital role in positively influencing its acceptance, especially if these are actors having a high social significance for the user, such as friends, family, work colleagues, or professional experts [111]. Arguments, for example applied to explain an AI and its results or actions [108], are accepted if they have the support that makes them acceptable to the participants in a conversation. Similarly, information that establishes perceived usefulness and ease of use is related to user acceptance [92]. The greater the coherence of a proposition with other propositions the greater its acceptability. If a proposition is highly coherent with a person's beliefs, then the person will believe the proposition with a high degree of confidence and the other way around, also known as confirmation bias [17, 101].

5.3 Information Representations and Interaction Methods

Methods for representing information and interacting with those in the context of human-AI collaboration are trusted if they exhibit a certain anthropomorphism such as natural language and speech [32, 27, 9, 105, 108, 42, 52, 83], text [17, 88] or human like visual appearances, for example in the context of robotics [38, 48, 92].

Textual and speech representations The former include text-based, as well as speech-based input and output. Especially domain experts expect system feedback in natural language to domain specific language [32] and be presented within 3-4 seconds [48, 69] to ensure a cognitive and emotional linkage through realistic, social interaction. Social quality of a dialogue through emotionally intelligent interaction can profit from closed-loop interaction with cognitive human models [48].In order to address the previously described information need by answering ”why” questions, but also “how” and “why not” questions, natural language should be expressed easily understandable in a narrative style [9].Dialectic explanations have for example been generated based on a log file with internal steps an AI performed to reach a certain recommendation [108].

Data and information visualizations In addition to natural language and speech interaction, data and information visualizations such as graphs [8, 115, 52], charts [79], and animations [17] are especially suitable for efficiently conveying information from statistical [78] and model-specific data [108, 42] such as intermediate network layers of DNNs [115], neuron activation and weights or token embedding in 2D and gradient based methods [57]. In addition, visualizations suitable for the representation of structural information such as CNN feature maps or DNNs graph structures [27, 57] or conceptual and semantic information [42, 57]. Features impact such as words impact on the classification outcome could effectively be represented through color-coding [115], especially when coding is based on relations relevant grammars [70]. Even though, speech is frequently used to represent explanatory information and multi-modal data contains persistent inconsistencies and biases [57], combining graphic narratives with natural language can be even more effective [9, 108, 118, 92, 79], reduce human workload or increase human performance [113]. Visual representations are particularly suitable for target groups with little background knowledge; analogies that correspond to the mental model can reinforce this [109].

Interaction quality Regardless of modality, safety-critical contexts often require interactive and reciprocal information exchanges and learning among humans and machines (HITL) [64, 57], while answers are expected to be fast and accurate [119]. Depending on user task and context touch and gesture-based interaction methods have also demonstrated to be powerful and effective [63] while emotion-aware mechanisms, especially when combined with human-like appearance support user satisfaction and adherence[92]. In any case suitability of a representation or interaction method is strongly dependent on age, culture and gender of the user group [54]. Examples for effective and efficient information representations and interaction methods include logged interactions to handle lost link procedures and error-free resumption of interaction after interrupted communication for an an artificial pilot. This system applied natural language interaction to interact with the terminal crew, to enable automated reasoning and decision making, to coordinate autonomous operations and basic pilot procedures with variable autonomy [69].

6 Discussion and Conclusion

This article aims to bridge the gap between technical and human perspectives in developing AI systems that are understandable, acceptable, and trustworthy. To achieve this, user needs are identified and transformed into requirements for AI system design, constituting an initial step for requirements engineering. These requirements must be validated and refined for various application domains to serve as the foundation for development activities.

6.1 Contribution

The results show that the existing methods for explaining AI (see Sec. 2.1) correspond to the needs and requirements of people with extensive background knowledge about AI, ML and DL models and whose task is to develop and improve AI models. In contrast, people who have little AI background knowledge use an AI system to achieve their individual goals and to process application-related tasks. They mainly expect the results and behavior of the system to be explained. Only occasionally the latter group would like to use the statistical parameters of an AI model to understand the system result or behavior. However, for this purpose, a lower and more flexible level of detail is required than for the former group. Relevant to either group yet is the questionable reliance of certain XAI methods, with many being prone to manipulation and adversarial attacks [95]. Arguably, multiple well-established explanation algorithms are criticized in [2] revealing that some fail to depict accurate mappings from input features to model outcomes [45, 62, 95]. To further provide explanations that fit the user needs, their requirements have to be taken into account during the development but also within AI applications (as described in Sec. 2.3). Since explanations should maximize the user's mental model of explanatory information, human feedback should be incorporated to a greater extent to iteratively improve development outcomes and AI result. As an example, such an approach was followed in [46] and appears highly promising.

A particular user group for AI systems collaborating with humans, are domain experts, such as ATCOs, medical professionals, or scientific personnel as crew and operators of space systems. They have extensive domain knowledge but restricted background knowledge of AI technologies. They also mainly need information that explains system results and behavior, aiming to understand a system outcome and behavior by information of the professional context rather than the technology. For medical professionals, this means, for example, that they want to interpret a result based on its relevance for different patient groups or based on its validity and relevance for other experts. Regarding the requirements for an AI system for air traffic control in terms of explainability, it can be stated that the needs of AI developers, as well as non-expert users, have to be considered: The former require information about the data and the model in terms of internal operations and the relationship between input and output variables. Essentially, the latter profit from information about the results and behavior of the AI system, why certain properties are assigned to specific results, and the contribution of features, rules, and decisions to a specific result. In general, the information should be presented in a contrasting manner and include various levels of detail depending on the context of each user group, the explanation capability of the AI, and the temporal, perceptual, and cognitive resources of individual users. For security-critical scenarios such as air traffic control, information should be communicated at conflict-free and less work-intensive times. Designing a comprehensible AI system requires various functionalities and modules that detail the individual needs and characteristics of all important user groups.

With respect to the information users require to accept an AI system, it can be stated that acceptance for AI systems profits from task-related information supporting users in achieving their goals, information demonstrating the performance and usefulness of the AI and information about privacy and ethical considerations. It has to be stated that information alone are not sufficient for acceptance the system in general needs to be useful in a user-friendly way also providing control over their data and data processing.

Regarding trustworthy information representation and interaction methods, results revealed that natural language and visual information representations are most suitable for human AI-collaboration, especially their combination. Effective interaction in safety-critical contexts, such as air traffic control, primarily require fast and accurate information exchange and learning between humans and machines. However, the suitability of a representation or interaction method is dependent on factors such as task and its context, the user's age, culture, and gender.

6.2 Limitations

Results presented describe user needs and requirements for a system only to the extent that these were included in the literature. In this context, the underlying data is subject to time-dependent biases towards technology-centered development methods, limited result validity, and biases due to the topicality of applied models. Early work, for example, developed models with much smaller data sets which is why the users' need to access these data might be more valid for earlier than for present systems. Literature analysis and synthesis was guided by the three research questions formulated in Sec. 3. In order to provide the most comprehensive and generally valid information possible, no restrictions were placed on the fields of application, user groups, technologies, or research methods/questions. Accordingly, considered papers exhibit a high level of heterogeneity regarding these characteristics. Nevertheless, contextual and methodological variance among studies examined must be acknowledged as a potential limitation. However, its effect is reduced by the fact that the user requirements formulated here are validated, refined, and supplemented for the air traffic control application context.

6.3 Future Work

This literature analysis demonstrated that, with respect to interdisciplinary perspectives in developing AI systems, different frameworks for considering humans are exploited which are not being integrated enough. On the one hand, the human-in-the-loop paradigm as a set of human oversight mechanisms on systems running AI models is well-known within the technical community. On the other hand, human factors specialists and psychologists have adopted a human-centered design approach, in a framework that develops socio-technical systems by involving the human perspective in all steps of the design process. Finally, in safety-critical contexts such as air traffic control, research and development focuses on shifting from manual control where the human is *in* the loop, to supervisory control where the human operator is *on* the loop. By having integrated automation in aviation some decades ago, the human operator no longer needs to be in direct control of the system. As a result operators are supervising many aspects of the system, which changes the role of the human in a system.

Therefore, one main topic of future work is to share and integrate the different perspectives and methods for designing understandable, acceptable, and trustworthy AI systems in an interdisciplinary development team. Another topic for further research lies on investigating and validating the presented findings for AI integration in air traffic control systems. A first step is to conduct user workshops as to assess their expectations on tasks to be allocated between human and AI system, on information needed from AI systems, user-friendly interaction and use of personal data. In doing so, a two-day workshop with nine ATCOs from German air navigation service provider (DFS) and Austro Control GmbH is currently being conducted. Furthermore, it is planned to research and validate prototypes of AI systems with users throughout the design process of AI systems in aviation. To achieve this, experimental studies will be conducted in laboratory settings simulating a control-center task environment, as well as large-scale simulations of air traffic control with experienced operators. In doing so, guidelines for effective and safe collaboration between AI systems and human operators in safety-critical contexts will be investigated, which will finally lead to recommendations for the development of AI systems.

Bibliography

[1] Explaining Trained Neural Networks with Semantic Web Technologies: First Steps (07/2017 2017), http://daselab.cs.wright.edu/nesy/NeSy17/

[2] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. p. 9525–9536. NIPS'18, Curran Associates Inc., Red Hook, NY, USA (2018)

[3] Ajzen, I.: The theory of planned behavior. Organizational behavior and human decision processes **50**(2), 179–211 (1991), https://doi.org/10.1016/0749-5978(91)90020-T

[4] Alshammari, M., Nasraoui, O., Sanders, S.: Mining semantic knowledge graphs to add explainability to black box recommender systems. IEEE Access **7**, 110563–110579 (2019). https://doi.org/10.1109/ACCESS.2019.2934633

[5] American Psychological Association and others: Apa dictionary of psychology online (2020)

[6] Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information fusion **58**, 82–115 (2020), https://doi.org/10.1016/j.inffus.2019.12.012

[7] Assael, Y., Sommerschield, T., Shillingford, B., Bordbar, M., Pavlopoulos, J., Chatzipanagiotou, M., Androutsopoulos, I., Prag, J., de Freitas, N.: Restoring and attributing ancient texts using deep neural networks. Nature **603**(7900), 280–283 (2022). https://doi.org/10.1038/s41586-022-04448-z

[8] Atkinson, D.J.: SHARP: Spacecraft health automated reasoning prototype. NASA. Johnson Space Center, Control Center Technology Conference Proceedings (Aug 1991), https://ntrs.nasa.gov/citations/19920002802

[9] Baclawski, K., Bennett, M., Berg-Cross, G., Fritzsche, D., Sharma, R., Singer, J., Sowa, J.F., Sriram, R.D., Underwood, M., Whitten, D.: Ontology summit 2019 communiqué: Explanations. Appl. Ontol. **15**(1), 91–107 (Feb 2020). https://doi.org/10.3233/ao-200226

[10] Bano, M., Zowghi, D.: Users' involvement in requirements engineering and system success. In: 2013 3rd International Workshop on Empirical Requirements Engineering (EmpiRE). pp. 24–31. IEEE (2013), https://doi.org/10.1109/EmpiRE.2013.6615212

[11] Beno, M.: Robot rights in the era of robolution and the acceptance of robots from the slovak citizen's perspective. In: 2019 IEEE International Symposium on Robotic and Sensors Environments (ROSE). pp. 1–7 (Jun 2019). https://doi.org/10.1109/ROSE.2019.8790429

[12] Beyret, B., Shafti, A., Faisal, A.A.: Dot-to-Dot: Explainable hierarchical reinforcement learning for robotic manipulation. In: 2019 IEEE/RSJ

International Conference on Intelligent Robots and Systems (IROS). pp. 5014–5019 (Nov 2019). https://doi.org/10.1109/IROS40897.2019.8968488

[13] Blöbaum, B., et al.: Trust and communication in a digitized world. Models and Concepts of Trust Research. Heidelberg et al.: Springer (2016), http://dx.doi.org/10.1007/978-3-319-28059-2

[14] Bonini, D.: Atc do i trust thee? referents of trust in air traffic control. In: CHI'01 Extended Abstracts on Human Factors in Computing Systems. pp. 449–450 (2001). https://doi.org/10.1145/634067.634327

[15] Braun, M., Bleher, H., Hummel, P.: A leap of faith: Is there a formula for "trustworthy" AI? Hastings Cent. Rep. **51**(3), 17–22 (May 2021), https://doi.org/10.1002/hast.1207

[16] Bruder, C., Jörn, L., Eißfeldt, H.: When pilots and air traffic controllers discuss their future (2008)

[17] Burkart, N., Huber, M.F.: A survey on the explainability of supervised machine learning. jair **70**, 245–317 (Jan 2021). https://doi.org/10.1613/jair.1.12228

[18] Cai, C.J., Winter, S., Steiner, D., Wilcox, L., Terry, M.: "hello ai": uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making. Proceedings of the ACM on Human-computer Interaction **3**(CSCW), 1–24 (2019), https://doi.org/10.1145/3359206

[19] Calvaresi, D., Mualla, Y., Najjar, A., Galland, S., Schumacher, M.: Explainable Multi-Agent systems through blockchain technology. In: Explainable, Transparent Autonomous Agents and Multi-Agent Systems. vol. 11763, pp. 41–58. Springer International Publishing (2019). https://doi.org/10.1007/978-3-030-30391-4_3

[20] Chen, Y.H., Chien, S.H., Wu, J.J., Tsai, P.Y.: Impact of signals and experience on trust and trusting behavior. Cyberpsychology, Behavior, and Social Networking **13**(5), 539–546 (2010), https://doi.org/10.1089/cyber.2009.0188

[21] Choung, H., David, P., Ross, A.: Trust in ai and its role in the acceptance of ai technologies. International Journal of Human–Computer Interaction pp. 1–13 (2022), https://doi.org/10.1080/10447318.2022.2050543

[22] Cooke, N.J., Gorman, J.C., Myers, C.W., Duran, J.L.: Interactive team cognition. Cognitive science **37**(2), 255–285 (2013), https://doi.org/10.1111/cogs.12009

[23] Cooke, N.J., Salas, E., Cannon-Bowers, J.A., Stout, R.J.: Measuring team knowledge. Human factors **42**(1), 151–173 (2000), https://doi.org/10.1518/001872000779656561

[24] Council, N.R., et al.: Research and modeling of supervisory control behavior: Report of a workshop (1930)

[25] Cui, Y., Koppol, P., Admoni, H., Niekum, S., Simmons, R., Steinfeld, A., Fitzgerald, T.: Understanding the relationship between interactions and outcomes in human-in-the-loop machine learning. In: Zhou, Z.H. (ed.) Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21. pp. 4382–4391. International Joint Conferences on Artificial Intelligence Organization (8 2021). https://doi.org/10.24963/ijcai.2021/599, survey Track

[26] Dalpiaz, F., Niu, N.: Requirements engineering in the days of artificial intelligence. IEEE software **37**(4), 7–10 (2020), https://doi.org/10.1109/MS.2020.2986047

[27] Dam, H.K., Tran, T., Ghose, A.: Explainable software analytics. In: Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results. pp. 53–56. ICSE-NIER '18, Association for Computing Machinery, New York, NY, USA (May 2018). https://doi.org/10.1145/3183399.3183424

[28] Davis, F.D.: A technology acceptance model for empirically testing new end-user information systems: Theory and results. Ph.D. thesis, Massachusetts Institute of Technology (1985), http://dspace.mit.edu/handle/1721.1/7582

[29] Day, D.: Application of AI principles to constraint management in intelligent user interfaces. In: Association for Information Systems, Proceeding of the Americas Conference on Information Systems. pp. 730–732 (1997), http://aisel.aisnet.org/amcis1997/54?utm_source=aisel.aisnet.org

[30] De, T., Giri, P., Mevawala, A., Nemani, R., Deo, A.: Explainable AI: A hybrid approach to generate Human-Interpretable explanation for deep learning prediction. In: Complex Adaptive Systems. vol. 168, pp. 40–48 (2020). https://doi.org/10.1016/j.procs.2020.02.255

[31] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848

[32] Dominick, W.D., Kavi, S.: Knowledge based systems: A preliminary survey of selected issues and techniques. Tech. Rep. DBMS.NASA/RECON-5 (May 1984), https://ntrs.nasa.gov/citations/19890005582

[33] Dunning, D., Fetchenhauer, D.: Understanding the psychology of trust. Psychology Press (2011)

[34] Díaz-Rodríguez, N., Lamas, A., Sanchez, J., Franchi, G., Donadello, I., Tabik, S., Filliat, D., Cruz, P., Montes, R., Herrera, F.: Explainable neural-symbolic learning (x-nesyl) methodology to fuse deep learning representations with expert knowledge graphs: The monumai cultural heritage use case. Information Fusion **79**, 58–83 (2022), https://doi.org/10.1016/j.inffus.2021.09.022

[35] Earle, T.C., Siegrist, M., Gutscher, H.: Trust, risk perception and the tcc model of cooperation. In: Trust in Risk Management, pp. 18–66. Routledge (2010)

[36] EASA: EASA concept paper: First usable guidance for level 1 machine learning applications (2021)

[37] Eder, K., Harper, C., Leonards, U.: Towards the safety of human-in-the-loop robotics: Challenges and opportunities for safety assurance of robotic co-workers'. In: The 23rd IEEE International Symposium on Robot and Human Interactive Communication. pp. 660–665 (2014). https://doi.org/10.1109/ROMAN.2014.6926328

[38] Ene, I., Pop, M.I., Nistoreanu, B.: Qualitative and quantitative analysis of consumers perception regarding anthropomorphic AI designs. In: Proceedings of the International Conference on Business Excellence. vol. 13, pp. 707–716 (2019). https://doi.org/10.2478/picbe-2019-0063

[39] European Commission, Directorate-General for Communications Networks, Content and Technology: The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment. Publications Office (2020). https://doi.org/10.2759/002360

[40] Finkelstein, A., Kramer, J.: Software engineering: a roadmap. In: Proceedings of the Conference on the Future of Software Engineering. pp. 3–22 (2000)

[41] Garibaldi, J.M.: The need for fuzzy AI. IEEE/CAA Journal of Automatica Sinica **6**(3), 610–622 (May 2019). https://doi.org/10.1109/JAS.2019.1911465

[42] Gaur, M., Faldu, K., Sheth, A.: Semantics of the Black-Box: Can knowledge graphs help make deep learning systems more interpretable and explainable? IEEE Internet Comput. **25**(1), 51–59 (Jan 2021). https://doi.org/10.1109/MIC.2020.3031769

[43] Gerdes, A.: The quest for explainable AI and the role of trust (work in progress paper). In: Proceedings of the European Conference on the impact of Artificial Intelligence and Robotics (ECIAIR). pp. 465–468 (2019), https://doi.org/10.34190/ECIAIR.19.046

[44] Gerlings, J., Shollo, A., Constantiou, I.: Reviewing the need for explainable artificial intelligence (xai). In: 54th Annual Hawaii International Conference on System Sciences, HICSS 2021. pp. 1284–1293. Hawaii International Conference on System Sciences (HICSS) (2021), https://doi.org/10.24251/HICSS.2021.156

[45] Ghorbani, A., Abid, A., Zou, J.: Interpretation of neural networks is fragile. Proceedings of the AAAI Conference on Artificial Intelligence **33**(01), 3681–3688 (Jul 2019). https://doi.org/10.1609/aaai.v33i01.33013681

[46] Ghorbani, A., Wexler, J., Zou, J.Y., Kim, B.: Towards automatic concept-based explanations. Advances in Neural Information Processing Systems **32** (2019)

[47] Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). pp. 80–89. ieeexplore.ieee.org (Oct 2018). https://doi.org/10.1109/DSAA.2018.00018

[48] Goodman, P.H., Zou, Q., Dascalu, S.M.: Framework and implications of virtual neurorobotics. Front. Neurosci. **2**(1), 123–129 (Jul 2008). https://doi.org/10.3389/neuro.01.007.2008

[49] Gorman, J.C., Cooke, N.J., Winner, J.L.: Measuring team situation awareness in decentralized command and control environments. In: Situational Awareness, pp. 183–196. Routledge (2017)

[50] Hale, J.L., Householder, B.J., Greene, K.L.: The theory of reasoned action. The persuasion handbook: Developments in theory and practice **14**(2002), 259–286 (2002), https://dx.doi.org/10.4135/9781412976046

[51] Hauland, G.: Measuring individual and team situation awareness during planning tasks in training of en route air traffic control. The International Journal of Aviation Psychology **18**(3), 290–304 (2008), https://doi.org/10.1080/10508410802168333

[52] Hepenstal, S., Zhang, L., Kodagoda, N., Wong, B.l.W.: Developing conversational agents for use in criminal investigations. ACM Trans. Interact. Intell. Syst. **11**(3-4), 1–35 (Sep 2021). https://doi.org/10.1145/3444369

[53] Ibrahim, A., Klesel, T., Zibaei, E., Kacianka, S., Pretschner, A.: Actual causality canvas: A general framework for Explanation-Based Socio-Technical constructs. In: ECAI 2020: 24th European Conference on Artificial Intelligence. vol. 325, pp. 2978–2985 (2020). https://doi.org/10.3233/FAIA200472

[54] Ismatullaev, U.V.U., Kim, S.H.: Review of the factors affecting acceptance of AI-Infused systems. Hum. Factors (Mar 2022). https://doi.org/10.1177/00187208211064707

[55] Iucci, A., Hata, A., Terra, A., Inam, R., Leite, I.: Explainable reinforcement learning for Human-Robot collaboration. In: 2021 20th International Conference on Advanced Robotics (ICAR). pp. 927–934 (Dec 2021). https://doi.org/10.1109/ICAR53236.2021.9659472

[56] Jentzsch, S.F., Hochgeschwender, N.: Don't forget your roots! using provenance data for transparent and explainable development of machine learning models. In: 2019 34th IEEE/ACM International Conference on Automated Software Engineering Workshop (ASEW). pp. 37–40. IEEE (2019)

[57] Joshi, G., Walambe, R., Kotecha, K.: A review on explainability in multimodal deep neural nets. IEEE Access **9**, 59800–59821 (2021). https://doi.org/10.1109/ACCESS.2021.3070212

[58] Kästner, L., Langer, M., Lazar, V., Schomäcker, A., Speith, T., Sterz, S.: On the relation of trust and explainability: Why to engineer for trustworthiness. In: 2021 IEEE 29th International Requirements Engineering Conference Workshops (REW). pp. 169–175. IEEE (2021)

[59] Kästner, L., Langer, M., Lazar, V., Schomäcker, A., Speith, T., Sterz, S.: On the relation of trust and explainability: Why to engineer for trustworthiness. In: 2021 IEEE 29th International Requirements Engineering Conference Workshops (REW). pp. 169–175 (Sep 2021). https://doi.org/10.1109/REW53955.2021.00031

[60] Kim, B., Khanna, R., Koyejo, O.O.: Examples are not enough, learn to criticize! criticism for interpretability. Advances in neural information processing systems **29** (2016)

[61] Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., sayres, R.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceed-

ings of Machine Learning Research, vol. 80, pp. 2668–2677. PMLR (10–15 Jul 2018), https://proceedings.mlr.press/v80/kim18d.html

[62] Kindermans, P.J., Hooker, S., Adebayo, J., Alber, M., Schütt, K.T., Dähne, S., Erhan, D., Kim, B.: The (Un)reliability of Saliency Methods, pp. 267–280. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-28954-6_14

[63] Klumpp, M., Hesenius, M., Meyer, O., Ruiner, C., Gruhn, V.: Production logistics and human-computer interaction—state-of-the-art, challenges and requirements for the future. Int. J. Adv. Manuf. Technol. **105**(9), 3691–3709 (Dec 2019). https://doi.org/10.1007/s00170-019-03785-0

[64] Kraiss, F.: Decision making and problem solving with computer assistance. Tech. Rep. NASA-TM-76008 (Jan 1980), https://ntrs.nasa.gov/citations/19800007713

[65] Krueger, F.: The Neurobiology of Trust. Cambridge University Press (2021)

[66] Kujala, S., Kauppinen, M., Lehtola, L., Kojo, T.: The role of user involvement in requirements quality and project success. In: 13th IEEE International Conference on Requirements Engineering (RE'05). pp. 75–84. IEEE (2005), https://doi.org/10.1109/RE.2005.72

[67] Kujala 1, S.: Effective user involvement in product development by improving the analysis of user needs. Behaviour & Information Technology **27**(6), 457–473 (2008), https://doi.org/10.1080/01449290601111051

[68] Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.R.: Unmasking clever hans predictors and assessing what machines really learn. Nature communications **10**(1), 1–8 (2019), https://doi.org/10.1038/s41467-019-08987-4

[69] Lowry, M., Bajwa, A., Pressburger, T., Sweet, A., Fry, C., Dalal, M., Schumann, J., Dahl, D., Karsai, G., Mahadevan, N.: Design considerations for a variable autonomy executive for UAS in the NAS. Tech. Rep. ARC-E-DAA-TN51256 (Jan 2018), https://ntrs.nasa.gov/citations/20180004247

[70] Lukyanenko, R., Castellanos, A., Storey, V.C., Castillo, A., Tremblay, M.C., Parsons, J.: Superimposition: Augmenting machine learning outputs with conceptual models for explainable AI. In: Advances in Conceptual Modeling. vol. 12584, pp. 26–34. Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-65847-2_3

[71] Maalej, W., Nayebi, M., Ruhe, G.: Data-driven requirements engineering-an update. In: 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP). pp. 289–290. IEEE (2019), https://doi.org/10.1109/ICSE-SEIP.2019.00041

[72] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D., Gebru, T.: Model cards for model reporting. In: Proceedings of the conference on fairness, accountability, and transparency. pp. 220–229 (2019), https://doi.org/10.1145/3287560.3287596

[73] Mitchell, M.: Why ai is harder than we think. In: Proceedings of the Genetic and Evolutionary Computation Conference. pp. 3–3 (2021), https://doi.org/10.1145/3449639.3465421

[74] Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., PRISMA Group*, t.: Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. Annals of internal medicine **151**(4), 264–269 (2009), https://doi.org/10.7326/0003-4819-151-4-200908180-00135

[75] Mohseni, S., Zarei, N., Ragan, E.D.: A multidisciplinary survey and framework for design and evaluation of explainable ai systems. ACM Transactions on Interactive Intelligent Systems (TiiS) **11**(3-4), 1–45 (2021), https://doi.org/10.1145/3387166

[76] Molnar, C.: Interpretable Machine Learning (2019), https://christophm.github.io/interpretable-ml-book/

[77] Montavon, G., Binder, A., Lapuschkin, S., Samek, W., Müller, K.R.: Layer-wise relevance propagation: an overview. Explainable AI: interpreting, explaining and visualizing deep learning pp. 193–209 (2019), https://doi.org/10.1007/978-3-030-28954-6_10

[78] Munzner, T.: Visualization analysis and design. CRC press (2014)

[79] Murphy, R.R.: Human-robot interaction in rescue robotics. IEEE Trans. Syst. Man Cybern. C Appl. Rev. **34**(2), 138–153 (May 2004). https://doi.org/10.1109/TSMCC.2004.826267

[80] Nuseibeh, B., Easterbrook, S.: Requirements engineering: a roadmap. In: Proceedings of the Conference on the Future of Software Engineering. pp. 35–46 (2000), https://doi.org/10.1145/336512.336523

[81] Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., et al.: The prisma 2020 statement: an updated guideline for reporting systematic reviews. International journal of surgery **88**, 105906 (2021), https://doi.org/10.1016/j.ijsu.2021.105906

[82] Papenfuss, A.: Phenotypes of teamwork–an exploratory study of tower controller teams. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. vol. 57, pp. 319–323. SAGE Publications Sage CA: Los Angeles, CA (2013), https://doi.org/10.1177/1541931213571070

[83] Pierrard, R., Poli, J.P., Hudelot, C.: Spatial relation learning for explainable image classification and annotation in critical applications. Artif. Intell. **292**, 103434 (Mar 2021). https://doi.org/10.1016/j.artint.2020.103434

[84] Prentzas, N., Nicolaides, A., Kyriacou, E., Kakas, A., Pattichis, C.: Integrating machine learning with symbolic reasoning to build an explainable AI model for stroke prediction. In: 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE). pp. 817–821 (Oct 2019). https://doi.org/10.1109/BIBE.2019.00152

[85] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer (2019). https://doi.org/10.48550/ARXIV.1910.10683

[86] Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., Prudden, R., Mandhane, A., Clark, A., Brock, A., Simonyan, K., Hadsell, R., Robinson, N., Clancy, E., Arribas, A., Mohamed, S.: Skilful precipitation nowcasting

using deep generative models of radar. Nature **597**(7878), 672–677 (Sep 2021). https://doi.org/10.1038/s41586-021-03854-z

[87] Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**, 533–536 (1986). https://doi.org/10.1038/323533a0

[88] Sachan, S., Yang, J.B., Xu, D.L., Benavides, D.E., Li, Y.: An explainable AI decision-support-system to automate loan underwriting. Expert Syst. Appl. **144**, 113100 (Apr 2020). https://doi.org/10.1016/j.eswa.2019.113100

[89] Salas, E., Cooke, N.J., Rosen, M.A.: On teams, teamwork, and team performance: Discoveries and developments. Human factors **50**(3), 540–547 (2008), https://doi.org/10.1518/001872008X288457

[90] Shafik, R., Wheeldon, A., Yakovlev, A.: Explainability and dependability analysis of learning automata based AI hardware. In: 2020 26th IEEE International Symposium on On-line Testing and Robust System Design (IOLTS) (2020), https://doi.org/10.1109/IOLTS50870.2020.9159725

[91] Sharma, C., Bhavsar, P., Srinivasan, B., Srinivasan, R.: Eye gaze movement studies of control room operators: A novel approach to improve process safety. Computers & Chemical Engineering **85**, 43–57 (2016), https://doi.org/10.1016/j.compchemeng.2015.09.012

[92] Shin, D.: Embodying algorithms, enactive artificial intelligence and the extended cognition: You can see as much as you know about algorithm. J. Inf. Sci. Eng. (Jan 2021). https://doi.org/10.1177/0165551520985495

[93] Silva, P.: Davis' technology acceptance model (tam)(1989). Information seeking behavior and technology adoption: Theories and trends pp. 205–219 (2015), http://dx.doi.org/10.4018/978-1-4666-8156-9.ch013

[94] Simpson, J., Kingston, J., Molony, N.: Internet-based decision support for evidence-based medicine. Knowledge-Based Systems **12**(5), 247–255 (Oct 1999). https://doi.org/10.1016/S0950-7051(99)00014-3

[95] Slack, D., Hilgard, S., Jia, E., Singh, S., Lakkaraju, H.: Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. pp. 180–186 (2020), https://doi.org/10.1145/3375627.3375830

[96] Sousa, P., Ramos, C.: A distributed architecture and negotiation protocol for scheduling in manufacturing systems. Comput. Ind. **38**(2), 103–113 (Mar 1999). https://doi.org/10.1016/S0166-3615(98)00112-2

[97] Spreeuwenberg, S.: Choose for AI and for explainability. In: On the Move to Meaningful Internet Systems: OTM 2019 Workshops. vol. 11878, pp. 3–8. Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-40907-4_1

[98] Suchman, L.: Centers of coordination: A case and some themes. Discourse, tools and reasoning: Essays on situated cognition pp. 41–62 (1997). https://doi.org/10.1007/978-3-662-03362-3_3

[99] Sutcliffe, A.: Scenario-based requirements analysis. Requirements Engineering Journal **3**(1), 48–65 (1998), https://doi.org/10.1007/BF02802920

[100] Taggart Jr, W., Tharp, M.O.: A survey of information requirements analysis techniques. ACM Computing Surveys (CSUR) **9**(4), 273–290 (1977), https://doi.org/10.1145/356707.356710

[101] Thagard, P.: Explanatory coherence. Behav. Brain Sci. **14**(4), 739–739 (1991). https://doi.org/10.1017/S0140525X00057046

[102] Theis, S., Schäfer, D., Bröhl, C., Schäfer, K., Rasche, P., Wille, M., Brandl, C., Jochems, N., Nitsch, V., Mertens, A.: Predicting technology usage by health information need of older adults: Implications for ehealth technology. Work **62**(3), 443–457 (2019). https://doi.org/10.3233/WOR-192878

[103] Theis, S., Schäfer, D., Schäfer, K., Rasche, P., Wille, M., Jochems, N., Mertens, A.: What do you need to know to stay healthy?–health information needs and seeking behaviour of older adults in germany. In: Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018) Volume V: Human Simulation and Virtual Environments, Work With Computing Systems (WWCS), Process Control 20. pp. 516–525. Springer (2019), https://doi.org/10.1007/978-3-319-96077-7_55

[104] Tomasello, M., Carpenter, M., Call, J., Behne, T., Moll, H.: In search of the uniquely human. Behavioral and brain sciences **28**(5), 721–735 (2005), https://doi.org/10.1017/S0140525X05540123

[105] Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., Pearson, G., Kaplan, L.: Rapid trust calibration through interpretable and Uncertainty-Aware AI. Patterns (N Y) **1**(4), 100049 (Jul 2020). https://doi.org/10.1016/j.patter.2020.100049

[106] Tran, P.N., Pham, D.T., Goh, S.K., Alam, S., Duong, V.: An interactive conflict solver for learning air traffic conflict resolutions. Journal of Aerospace Information Systems **17**(6), 271–277 (2020). https://doi.org/10.2514/1.I010807

[107] Umbrello, S., Yampolskiy, R.V.: Designing AI for explainability and verifiability: A value sensitive design approach to avoid artificial stupidity in autonomous vehicles. International Journal of Social Robotics **14**(2), 313–322 (Mar 2022). https://doi.org/10.1007/s12369-021-00790-w

[108] Vassiliades, A., Bassiliades, N., Patkos, T.: Argumentation and explainable artificial intelligence: a survey. Knowl. Eng. Rev. **36**, e5 (2021). https://doi.org/10.1017/S0269888921000011

[109] Veitch, E., Alsos, O.A.: Human-Centered explainable artificial intelligence for marine autonomous surface vehicles. J. Mar. Sci. Eng. **9**(11), 1227 (Nov 2021). https://doi.org/10.3390/jmse9111227

[110] Verma, S., Arthur, A., Dickerson, J., Hines, K.: Counterfactual explanations for machine learning: A review https://arxiv.org/abs/2010.10596

[111] Vorm, E.S.: Assessing demand for transparency in intelligent systems using machine learning. In: 2018 Innovations in Intelligent Systems and Applications (INISTA). pp. 1–7 (Jul 2018). https://doi.org/10.1109/INISTA.2018.8466328

[112] Wickens, C., Mavor, .A., McGee, J.E.: Flight to the future: Humans factors in air traffic control (1997)

[113] Wickens, C.D., Helton, W.S., Hollands, J.G., Banbury, S.: Engineering psychology and human performance. Routledge (2021), https://www.routledge.com/Engineering-Psychology-and-Human-Performance/Wickens-Helton-Hollands-Banbury/p/book/9781032011738

[114] Wilson, T.D.: On user studies and information needs. Journal of documentation **37**(1), 3–15 (1981)

[115] Winkler, J.P., Vogelsang, A.: “what does my classifier learn?” a visual approach to understanding natural language text classifiers. In: Natural Language Processing and Information Systems. vol. 10260, pp. 468–479. Springer International Publishing (2017). https://doi.org/10.1007/978-3-319-59569-6_55

[116] Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., He, L.: A survey of human-in-the-loop for machine learning. Future Generation Computer Systems **135**, 364–381 (2022). https://doi.org/10.1016/j.future.2022.05.014

[117] Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., Tenenbaum, J.B.: Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. In: Advances in Neural Information Processing Systems (NIPS) (2018). https://doi.org/10.48550/ARXIV.1810.02338

[118] Yokoi, R., Nakayachi, K.: Trust in autonomous cars: Exploring the role of shared moral values, reasoning, and emotion in Safety-Critical decisions. Hum. Factors **63**(8), 1465–1484 (Dec 2021). https://doi.org/10.1177/0018720820933041

[119] Zarka, R., Cordier, A., Egyed-Zsigmond, E., Lamontagne, L., Mille, A.: Trace-based contextual recommendations. Expert Syst. Appl. **64**, 194–207 (Dec 2016). https://doi.org/10.1016/j.eswa.2016.07.035

[120] Zheng, M., Zhang, S., Zhang, Y., Hu, B.: Construct food safety traceability system for people’s health under the internet of things and big data. IEEE Access **9**, 70571–70583 (2021). https://doi.org/10.1109/ACCESS.2021.3078536