A Taxonomy of Explanation Types and Need Indicators in Human–Agent Collaborations

Lennart Wachowiak[1*], Andrew Coles[1], Gerard Canal[1†], Oya Celiktutan[2†]

[1*]Department of Informatics, King's College London, London, United Kingdom.
[2]Department of Engineering, King's College London, London, United Kingdom.

*Corresponding author(s). E-mail(s): lennart.wachowiak@gmail.com;
Contributing authors: andrew.coles@kcl.ac.uk; gerard.canal@kcl.ac.uk;
oya.celiktutan@kcl.ac.uk;
†Equal senior contribution.

Abstract

In recent years, explanations have become a pressing matter in AI research. This development was caused by the increased use of black-box models and a realization of the importance of trustworthy AI. In particular, explanations are necessary for human–agent interactions to ensure that the user can trust the agent and that collaborations are effective. Human–agent interactions are complex social scenarios involving a user, an autonomous agent, and an environment or task with its own distinct properties. Thus, such interactions require a wide variety of explanations, which are not covered by the methods of a single AI discipline, such as computer vision or natural language processing. In this paper, we map out what types of explanations are important for human–agent interactions, surveying the field via a scoping review. In addition to the typical introspective explanation tackled by explainability researchers, we look at assistive explanations, aiming to support the user with their task. Secondly, we survey what causes the need for an explanation in the first place. We identify a variety of human–agent interaction-specific causes and categorize them by whether they are centered on the agent's behavior, the user's mental state, or an external entity. Our overview aims to guide robotics practitioners in designing agents with more comprehensive explanation-related capacities, considering different explanation types and the concrete times when explanations should be given.

Keywords: explainable robotics, XAI, human–robot interaction, collaboration, survey, scoping review

1 Introduction

While the ability to explain is crucial in human–agent interactions and robotics [1–5], the field is missing a comprehensive overview of situations in which explanations might become necessary. Most of the current explanation generation methods focus on explaining the agent's inner workings, thus providing a form of transparency. However, human–agent collaborations require explanations that go beyond those solely taking into account the agent's inner workings. Firstly, the agent's human collaborator plays a crucial part in achieving the team's goals and might encounter problems on their own that lead to the need for an assistive explanation. In addition, making agents and humans work in a shared environment with common goals will lead to the emergence of social situations, including conversations with the agent,

that encompass yet another set of explanation types. This diversity of explanations relevant to social robotics is underexplored.

In this paper, we map and structure the landscape of explainability research in human–agent interactions and collaborations. The current lack of a clear understanding of the contexts in human–agent interactions that require explanations leads us to the following research questions:

- What can an agent explain in human–agent collaboration?
- What causes the need for an explanation in these situations?

We answer these questions by providing a comprehensive classification of: (1) explanation types, such as vision or joint movement explanations, and (2) causes of the need for an explanation, such as agent errors or uncertainties. We explicitly do not focus on explainability methods themselves, which can vary depending on the context (e.g., language generation vs. pathfinding explanations) and the underlying model (e.g., neural model vs. AI planner), and are covered in respective overviews referenced throughout. Our taxonomies were developed in a top-down manner, after which they were validated and refined through a PRISMA scoping review of the field. The created taxonomies contextualize the concept of explanations within the field of human–agent interaction and robotics. Therefore, our paper can help robotics researchers identify what explanation abilities might be required when a developed robot comes in contact with human supervisors or co-workers. Moreover, knowing what causes the need for an explanation will help with developing agents that can give explanations at the right time in a proactive manner — an important skill in real-time collaborations.

2 Background

2.1 What is an Explanation?

In order to grasp what contexts require explanations, we first need to define the term explanation. In science, separate fields, such as philosophy of science, social psychology, and cognitive psychology, emphasize different aspects of explanations and might even define the term differently [18]. However, looking at dictionary definitions can already provide us with insights. According to the Cambridge Dictionary [19], an explanation is defined as "the details or reasons that someone gives to make something clear or easy to understand". According to Merriam Webster [20] an explanation is "the act or process of explaining" with explaining being defined as: (1) "to make known", (2) "to make plain or understandable", (3) "to give the reason for or cause of", (4) "to show the logical development or relationships of". From these definitions, we can gather that at the core of an explanation is the provision of some information that leads the explanation's recipient to understand or know something. However, for an explanation, not every type of information suffices. For example, providing the time when asked "What time is it?" would not count as an explanation — it is a simple fact, unconnected to a larger body of knowledge. Thus, many researchers only regard something as an explanation if it refers to causes [11], which, however, is a contested view [21]. To keep the definition used in this paper as inclusive as possible and cover diverse human–agent interactions, we will adopt Faye's pragmatic-rhetorical account of explanations [22]. Faye defines an explanation as an answer to an explanation-seeking question, that is, a question that poses an epistemic problem. To resolve the problem of not accepting simple facts like the current time as an explanation, Faye further requires that "the question is answered with reference to other facts" and that "this connection, by being brought to the questioner's attention, improves her understanding of the fact mentioned in the question".

2.2 Existing Explanation Classifications

A concern for explanations can be found in various subfields of AI, from computer vision to natural language processing and robotics. The motivations, the ways to evaluate, and the methods to generate explanations vary widely. This diversity has been argued to be a good thing as it leads to creativity and novel approaches being developed instead of narrowing down the field too much too early [23]. At the same time, it means that in robotics, a field that combines various AI techniques, there are many potential

Table 1 Ways to categorize explanations found in literature (based on generation and communication of explanation)

Categorization	Description
Generation of explanation	
Ante-hoc vs. Post-hoc	Ante-hoc systems are transparent by nature, e.g., decision trees. Post-hoc systems generate additional explanations to make their decision-making transparent [4, 6, 7].
Model-specific vs. model-agnostic	The XAI method can be designed for the logic of a specific model or work independently from the underlying logic [7].
Local vs. global	The explanation can target an individual, local decision of the agent or target the global behavior, that is, try to explain how the agent makes decisions in general [7].
Communication of explanation	
Modality of presentation	Explanations can be presented in various ways, for example, as text, visualization, or expressive motion [2].
Proactively vs. on-request	The explanation can be given proactively by the agent or after it was requested by the user.
Before vs. during vs. after execution	An explanation can be given before an agent executes its action, during execution, or after execution [1]. Depending on the timing, its purpose might differ, for example, asking for approval before execution vs. providing a retrospective account [4, 8].
Static vs. interactive	An explanation can be provided as a static object or be presented in the form of an interactive dialogue in which the user can request additional information [6].

approaches and questions to consider. In the following, we will summarize typical classification schemes for explanations, e.g., model-agnostic vs. model-specific explanations. We collected these existing schemes from surveys on explainable AI (XAI), especially those with a focus on explanations in human–agent interaction and robotics [1–7, 24], as well as social science and philosophy papers concerned with defining explanations [11, 18, 25]. As a result, Tables 1 and 2 show how explanations are typically classified and provide references for more detailed descriptions of the respective scheme. We group the identified schemes into those distinguishing between: (1) generation methods, (2) ways of communication, (3) explanation recipients, and (4) content types.

In contrast to existing classification schemes, our proposed taxonomy does focus on the contextual situations in which an explanation becomes necessary. We identify and hierarchically organize such high-level situations and areas in the field of human–agent interaction by showcasing *what* can be explained and *when* a need for explanations arises.

2.3 Causes of the Need for an Explanation

While existing explainability surveys do not tackle the question of what causes the need for an explanation during a human–agent interaction, many surveys summarize the potential benefits of an explanation. These benefits are a good starting point for our investigation as they give hints at the reasons why an explanation is given — in some cases, for example, an explanation might be given due to the desire to reap a potential benefit. Miller [11] provides an overview of why AI systems might provide explanations. Although including some of the causes highlighted in our paper, he mostly focuses on why the user might want to receive an explanation and how it allows them to learn while also summarizing some of the social functions, such as persuasion. Keil gives an overview of what explanations are for, without an AI-setting in mind [26]. The purposes include being able to better predict future events, diagnose errors, attribute blame, justify an action, and derive increased aesthetic pleasure.

Similar to us but not in an AI context, Liquin and Lombrozo look at what causes the need

Table 2 Ways to categorize explanations found in literature (based on target and content of explanation)

Categorization	Description
	Target of explanation
Regular user vs. expert vs. external entity	Explanations can be geared towards different entities: the user that the agent interacts with, the developer of the agent who wants to improve or debug it, or an external entity, e.g., policymakers regulating how explainable a system needs to be [1].
User-aware vs. non-user-aware	XAI research often ignores the user who has to engage with the generated explanations. In contrast, user-aware explanations take into account how the user perceives the world and how an explanation might improve their understanding [2, 9]. User-aware explanations can also take into account users' preferences, for instance, based on cultural backgrounds [10].
	Content of explanation
Why vs. how vs. what	Explanations try to answer an explicitly or implicitly voiced question from the user. The potential questions are often grouped into how, what, and why the agent is doing something, with why questions being the most common ones [3, 11].
Marr's levels of explanations	Marr [12] differentiates between three levels at which a system can be understood: the computational level (what is the goal of a computation), the algorithmic level (how is the computation implemented), and the hardware level (on what substrate is the computation realized).
Deductive proof vs. causal pattern vs. mental model	Different streams of thought consider explanations to be different things, e.g., deductive proofs as in logic, a disclosure of causal relations, or patterns of neural activation [13].
Intentional vs. non-intentional	Folk psychology differentiates between intentional and non-intentional behavior and the explanations given for both [14].
Folk explanation coding scheme	Malle's coding scheme [15] distinguishes four main types of explanations. Unintentional behavior is explained with cause explanations, while intentional behavior is usually explained with reason explanations or causal history of reasons. Lastly, there is the group of explanations that refers to the factors that enabled an action. Each category can be further subdivided, for example, intentional acts can be explained with regard to the agent's beliefs, desires, or values.
Internal vs. external factors	An explanation can attribute an event or action to external factors, such as the environment and situation, or internal factors, such as the traits of the person who performed the action [14].
Argument schemes	Agents can provide an explanation in the form of an argument for their decision [16]. The field of argumentation provides rich literature on different types of arguments (e.g., [17]).
Contrastive vs. non-contrastive	While non-contrastive explanations explain an event in isolation, contrastive explanations contrast the event that needs an explanation to another event that the recipient of the explanation expected instead [11]. Thus, contrastive explanations answer questions such as "Why did you do A instead of B"?
Agent-centric vs. user-centric	The explanation can be about the agent's decision-making (agent-centric) or about persuading the user to take a certain action as in the case of recommender systems (user-centric) [1].

for an explanation [27]. They empirically test a set of candidate indicators of the need for an explanation. To do so, participants were given why-questions and were asked how strongly these demand explanations. Additionally, participants had to rate 13 potential determinants of the need for an explanation, for example, the expected future utility of receiving the explanation. Besides expected utility, requiring expertise to answer the question was a strong predictor of the need for an

explanation. Other valuable predictors included the user's prior knowledge of the topic and the explanation's expected information content.

In addition to underlying reasons for why an explanation is given and its benefits, one can also look at the immediate triggers that lead to an explanation being provided. Krause and Vossen provide a summary of such triggers, differentiating between direct and indirect ones. While direct triggers include the human asking a question or giving a command, indirect triggers include the agent being uncertain or detecting the human being confused [28]. Later in the paper, we will distinguish between such immediate triggers and the underlying causes of the need for an explanation.

Compared to the majority of literature, our overview focuses on what initially causes the need for an explanation and not on the general benefits of explanations. The resulting compilation of causes was specifically created with human–agent interactions in mind, leading to a unique overview with particular domain-specific causes being identified.

3 Method

We created the taxonomies presented in Section 4 and Section 5 through (1) discussion among experts and reference to selected literature and (2) validation and refinement of the initial drafts via a scoping literature review. The first phase included discussions among the four authors of the paper, all of them being robotics researchers, and the consultation of selected papers relevant to our ideas, thus iteratively drafting the taxonomy in a top-down fashion. The second phase was used to validate this initial draft and find blind spots by systematically consulting the literature. This bottom-up analysis followed the PRISMA guidelines for a scoping literature review [29], which provides a clear reporting checklist.

For the scoping review, we searched for papers in the ACM Digital Library, IEEE Xplore, and Scopus. In Scopus, the subject areas were limited to computer science and engineering. Our eligibility criteria were that the paper:

- is published between January 2013 and June 2023,
- is written in English,
- is published in conference proceedings, workshop proceedings, or a journal,
- is not a survey, position paper, or project proposal.
- refers to an explainability component for an agent or a robot.

In regards to the last point, we used the broad understanding of the term "explanation" as defined in Section 2.1. This allowed us to consider all contexts in which explanations are used and not just find papers that are aligned with the most common notion or preconceived ideas of explainability. Importantly, the explanation is required to come from a robot or agent. Explanations given by users or simple apps, e.g., a museum app that explains the exhibits, are not included.

We used the following keywords to find papers:

- explainable AND robot
- OR explainable AND robotics
- OR "explainable agent".

Using the term "agent" in addition to "robot" allowed us to also capture the use of explanations in simulations and games where many interaction scenarios are tested first. Keywords were checked against paper titles, abstracts, and keyword lists. The term "explainable agent" was required to appear in this exact manner as, otherwise, results included too many unrelated papers. As an example, this resulted in the search string *("Document Title":explainable robot) OR ("Document Title":explainable robotics) OR ("Document Title":"explainable agent") OR ("Abstract":explainable robot) OR ("Abstract":explainable robotics) OR ("Abstract":"explainable agent") OR ("Index Terms":explainable robotics) OR ("Index Terms":"explainable agent") OR ("Index Terms":explainable robotics)* for IEEE.

The resulting 526 papers were organized in a CSV, allowing us to remove duplicates. Afterwards, we screened the titles and abstracts, excluding papers that did not meet our eligibility criteria. Lastly, we did a final eligibility check, consulting relevant parts of the full papers, and then classified the included 227 papers based on our taxonomy drafts created in phase one. In the case a paper did not fit into one of our categories, we improved the taxonomy to reflect that paper's content. Notably, a paper was only considered

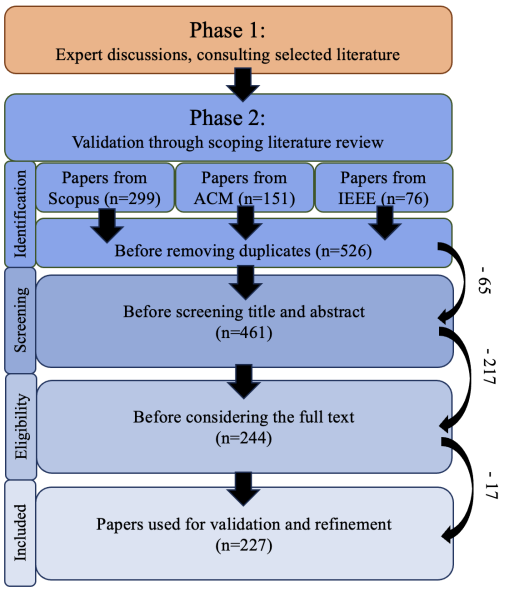

Fig. 1 Taxonomy creation procedure

to contain a cause of the need for an explanation if it used its explanations in a user study. Otherwise, it was too vague how the researchers planned to use the explanation during a human–agent interaction, making it unclear what would cause the explanation need. Figure 1 summarizes the approach. A final list of included papers and the respective classification decisions are available online: https://github.com/lwachowiak/Explanation-Types-and-Need-Indicators-in-HAI. The resulting taxonomies are presented in the following sections.

4 Taxonomy of Explanation Types

In recent years, XAI research has increasingly focused on making complex black-box models more interpretable [7]. However, this narrow view of explainability excludes many types of explanations applicable to AI systems [30]. In human–agent collaborations, it is not only the agent's decision-making that can warrant an explanation, as the collaboration is embedded in a task and environment with distinct rules and goals. Therefore, the task and environment themselves might cause the user to be confused and require an assistive explanation to help them progress. We can, thus, see that the ability to explain requires much more from cooperative agents than just making their own decisions transparent. Figure 2 presents a literature-based taxonomy of different types of explanations that are potentially required in the context of human—agent collaborations. At the top level, it is divided into two main branches, *assistive* and *introspective explanations*. Assistive explanations are given to users who need help as they do not understand something related to the task or environment, whereas introspective explanations are given to explain the agent's behavior. Further sub-levels cover different robot modules and areas of interaction, which can sometimes overlap when generating an explanation in practice.

4.1 Assistive Explanations

We subdivide assistive explanations into *task-oriented assistive explanations* and *general knowledge explanations*. Task-oriented explanations are about helping the user in situations where they are stuck due to missing knowledge about the task and environment (e.g., [31]) or due to high task complexity that does not allow them to quickly identify the best action (e.g., [32]). An example of a question coming from the user that demands a task-oriented assistive explanation is "What do I have to do next?". Having a mental model of the interaction partner and knowing what and how they want to achieve a goal is one of the key challenges when giving good task-oriented assistive explanations [33], as this allows the explanation to target missing or wrong knowledge. Furthermore, during collaborations, the user and agent might disagree on the best course of action, thus, leading to the need for introspective explanations justifying the assistive explanation from the agent for the user.

The second type of assistive explanation provides the user with a piece of general knowledge that is not related to a specific task. For example, the user might ask their personal robotic assistant, "Why is the sky blue?" which is unrelated to a shared goal-oriented task of the agent and the human but simply arises due to the user's intrinsic curiosity. In this case, the robot can provide a scientific explanation that is neither about itself nor a task, which might not even exist in this context. Such explanations also play a role when the user wants to learn something about a specific topic,

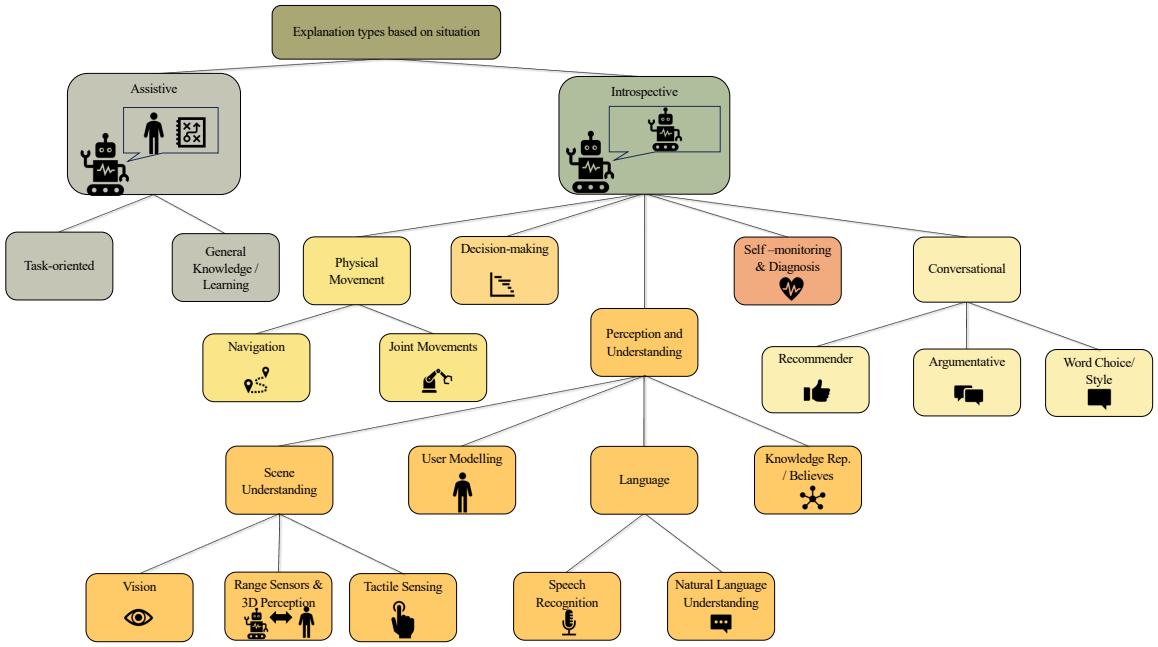

Fig. 2 Taxonomy of Explanation Types.

which has been explored in research, for instance, through robot guides in museums [34] or robots facilitating language learning [35].

4.2 Introspective Explanations

XAI research typically focuses on introspective explanations as these are the explanations targeting the issue of black-box models, which became prominent with the rise of deep neural networks [36]. Neural models compute their output through a usually large number of computations of millions or even billions of neurons chained together across multiple layers — a characteristic that makes it difficult to comprehend why a specific output was given. However, neural models are not the only type of algorithm that requires explanations in human–robot interaction, but similar XAI research is conducted for motion planners or agents selecting their actions through planning or behavior trees. In the following, we present different high-level types of explanations or explanation contexts that are relevant in robotics independent of their underlying implementation. On the top level of introspective explanations, we distinguish between *decision-making*, *physical movement*, *perception and understanding*, *self-monitoring*, and *conversational explanations*.

Decision-making

Explanation can be about the agent's high-level *decision-making*, answering a question such as "Why did you decide to do action A instead of action B?". The explanation with which the agent answers this question depends on the underlying architecture, as fields such as explainable planning [37] and explainable reinforcement learning [38] provide different explanations and methods to generate them.

Physical Movement

Another type of task-related introspective explanation can target the agent's *physical movement*. Firstly, such explanations might target an agent's *navigation* [39] capabilities, answering a question such as "Why did you take this (suboptimal) path?". Secondly, such movement explanations might target the motion planning for *joints* [40], thus answering a question such as "Why did you grasp the object like that?". The line between these two sub-areas can get blurry when considering the movement of joints that allow a robot to advance over a terrain.

Perception and Understanding

Another area in need of explanations is the agent's *perception and understanding* of the world, task,

and collaborator. Explanations about an agent's scene understanding might answer questions such as "Why did you think this is an apple?". An explanation to this question can refer to the agent's *visual processing* [41] or object recognition via *range sensors* [42]. Sometimes, even *tactile sensors* play a role that can be explained [43]. Besides understanding the scene itself, in human–agent interactions, it is also crucial for an agent to understand their human collaborator. These explanations are still introspective as they are about the agent's understanding of its collaborator and not about assisting the user with solving an external problem as before. Various aspects of the *agent's mental model of the user* can be explained, for instance, how the agent predicts the user to move [44] or why the agent assigns a user a specific affective state [45]. Moreover, in interacting with a human, the agent has to perceive and understand language. Explanations can target the robot's *speech recognition* [46] or its *natural language understanding* [47], e.g., why it started pursuing a specific goal based on a language command. Lastly, explanations can target an agent's more abstract *knowledge representation*, showcasing why it holds certain beliefs or at least making the beliefs transparent [48]. In comparison to the decision-making explanation, such belief explanations do not have to be action-oriented.

Self-monitoring

Self-monitoring explanations shed light on a robot's comprehension of its own state, that is, the state of software and hardware modules, including potential malfunctioning [49, 50]. An example of an explanation-demanding question is "Why is your battery so low?" or "Why is this module in a failure mode?". Answers to these questions could refer to broken hardware parts or previously undertaken activities.

Conversational

Conversational explanations are given in the context of a conversation between an agent and a human. *Argumentative explanations* are given during a discussion and might be triggered by a question such as "Why do you think this law is bad?". The argumentative explanation can then explain the agent's beliefs or arguments. Furthermore, *recommender explanations* are given if the agent functions as a recommender system, for instance, giving the human ideas for what movie to watch or how to be healthy [51, 52]. Lastly, explanations can also be given regarding specific *word choices* [53].

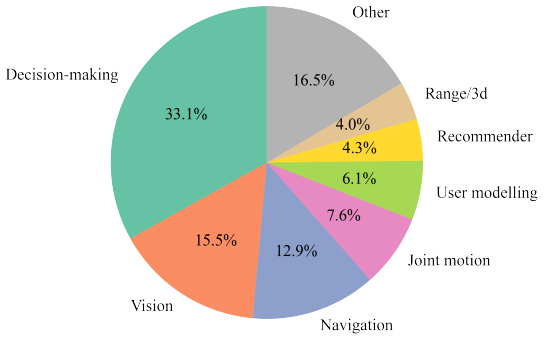

Fig. 3 The most common explanation types found in literature

4.3 Frequency in Literature

The scoping review not only allowed us to verify and improve our initial taxonomy but also gave us an impression of how frequently each explanation type is considered in the literature. Figure 3 shows that, in research papers, the explanations of robots or agents most commonly target the agent's decision-making (33.1% of all labels), vision capabilities (15.5%), and navigation (12.9%). The remaining 38.5% comprises the rest of our taxonomy's categories; however, each category contributes less than 10%. It should be noted here that the percentage value corresponds to the share of all labels and not papers, as a single paper can have multiple labels corresponding to different types of explanations. This can be the case when two separate explainability modules are tackled in the same paper, as is the case with 44 of the reviewed papers. Secondly, multiple labels can occur when a single explanation module covers multiple of the defined categories at the same time. For instance, the explanation of a navigation decision might involve not only a map of the area but also up-to-date scene understanding from vision or range sensors, potentially leading to joint explanations. Similarly, conversational choices made by the agent might be explained with reference to the agent's language understanding in regard to what the conversational partner said.

When considering these numbers, it is important to keep in mind the keywords that were used to identify relevant papers (explainable + robot, explainable + robotics, explainable agent). For each category, like explainable vision, many more papers could potentially be found, however, likely with little or no focus on robots or agents.

5 Causes of the need for an explanation

Why does the need for an explanation arise in human–agent collaborations? Section 5.1 and Table 3 summarize the causes that make an explanation necessary. These results were obtained through the literature review methodology described in Section 3. Of the 227 full-text papers we considered, 70 contained a user study from which we could infer what caused the need for an explanation to arise. Section 6 describes the relationship between this classification and the one presented in the previous section.

5.1 Cause Classification

Agent-centered Causes

The first group of causes is centered on what the agent does. This group of agent-centered causes includes cases in which an action taken by the agent goes wrong, e.g., the agent making an *error* or *violating a social norm*. Explanations can help with identifying the issue or serve as a form of justification or apology for what happened. Moreover, explanations can be helpful during action selection, when the agent might be *uncertain* about what to do or realize its *inability* to fulfill a request. An explanation can then highlight the issue to the user, who might respond by changing the goal or by helping the agent. During task execution, the agent might also encounter *unforeseen circumstances*, making its current course of action look weird to an external observer. An explanation highlights to the user why the agent suddenly needs to adapt its plan or can be a sign for the user to update the agent's world model. Lastly, the agent might give explanations with a specific *social purpose*, such as to convince or deceive the explanation recipient.

User-centered Causes

The second group of causes is centered on the user and their mental model. This group of user-centered causes includes the user having an *incomplete mental model* of the agent or task, a *mismatch of the user's mental model with the agent's model*, or *high task complexity*. Explanations help by providing those pieces missing in the user's mental model or can start a conversation about whose world model is the correct one. Lastly, explanations can again be given for a *social purpose*; however, this time, due to the social needs of the user, such as unwarranted distrust in the agent's decision or the need for interaction.

External Causes

Lastly, the need for an explanation can arise due to causes that are external to the direct human–agent interaction. These *external causes* include laws or norms requiring AIs to explain their behavior or programmers wanting to improve an agent or find bugs by better understanding an implementation.

5.2 Frequency in Literature

As in Section 4.3, we can observe the frequency of discussed causes of the need for an explanation in literature. However, the numbers presented here might show a somewhat skewed distribution of explanation causes. This is because we only included papers with user studies for this section as those were more concrete in when explanations would be provided, and still, many papers were vague on what causes an explanation. Of the 70 papers with user studies, more than half discussed social reasons for giving explanations, specifically, explanations given to increase trust and perception of the agent by the user (36 papers). Second-most commonly, explanations are given due to explicitly mismatched mental models between agent and user or users' incomplete mental models of task or agent (23 papers). The third biggest category is agent errors (8 papers), with other categories being represented by 5 or fewer papers.

Table 3 Reasons for requiring explanations

Cause	Definition	Example
Agent-centered causes of need for explanation		
Agent error	An agent making an error is a common scenario used in literature to study user reactions, e.g., changes in confusion and trust [54–56]. Explanations are a potential recovery strategy for the agent after an error occurs [57, 58].	The agent grasps the wrong object, takes a sub-optimal path, or becomes unresponsive due to software issues.
Agent ability	The agent is unable to do what is expected of it [59, 60] as it does not have the required abilities or permissions.	An object is out of reach, its shape is too difficult to grab, or it is too large or heavy to carry.
Unforeseen circumstances	The agent encounters the world to be in a state different from its world model and needs to adapt [61].	The agent first goes to the fridge to grab something, then realizes it is not there and has to re-plan.
Uncertainty	The agent is uncertain which action will lead to the best results. An explanation allows the user to correct the course of action or trust the agent's decision [28].	The agent assumes the object it needs is in Room A. But, it also assigns a non-zero probability for the object to be in Room B. To prevent a potential mistake and confusion, it tells the user why it is going to A.
Social norm violation	The agent does not adhere to the social script [62–64]. Its behavior deviates from what is expected in a social situation, e.g., through impoliteness or delays.	The agent interrupts the user or asks for something they already had asked for.
Social purpose	An agent can give an explanation with a social purpose, for example, to convince or to deceive the human [14, 65], or to improve how it is perceived.	The agent gives an explanation to persuade someone to change their beliefs or actions.
User-centered causes of need for explanation		
Incomplete user mental model of the agent or task	The user does not understand the agent's behavior, task, or environment as some piece of knowledge about them is missing from the user's mental model [9, 31]. Filling in those missing pieces via explanations helps with solving the task and predicting and understanding agent behavior, thus generating trust [18].	The user does not understand the task goals, what preconditions need to be fulfilled to execute some action, or where an object is.
Mismatch of the user's and the agent's mental model	User and agent model differ in regards to the environment, task, or each other. An explanation helps reconcile these differences [66].	The user thinks there is a better, more efficient order of actions than the one suggested by the agent.
High task complexity	The scenario is complex, and the user does not immediately understand how to solve the problem [32]. An explanation helps the user solve the task.	The task requires the user to come up with a complicated strategy, or the task requires the user to understand complicated rules.
Social purpose	Explanations can be given due to reasons concerning the social relationship between the agent and user. For instance, the user might have a desire for social interaction with the agent, leading to explanation-demanding questions [14]. Alternatively, the user might be scared of the agent or distrust it [67], which often goes hand in hand with an incomplete model of the agent.	The user asks the agent for an explanation of something just to make conversation or because they want to get to know the agent better.
External causes of need for explanation		
Required justification due to law	It is tested if the agent abides by policies and laws [7]. Here, the explanation need does not emerge from the interaction but arises due to external pressure.	It is checked if an agent adheres to explainability guidelines as outlined in regulations like GDPR.
Debugging and improvement of agent	The developer of an agent might use explanations to debug and improve the agent's behavior [7]. This can help identify errors, biases, or even cyber-attacks [68].	The programmer uses XAI methods for an agent's vision component as it often confuses two objects.

6 Relation between Types, Causes, and Triggers

Explanation Causes and Types

The causes mentioned in Section 5.1 are what make explanations in human–agent interactions needed. A specific cause, e.g., to convince, can be relevant for each of the explanation types from Section 4. For instance, an agent wanting to convince the user of a particular point of view might give an assistive explanation to make the user realize that they should do something in a certain way. Similarly, an agent might want to convince the user that the path it took was optimal, thus giving an introspective path planning explanation. However, other causes do not apply to all explanation types: an agent error often triggers an introspective explanation about why the agent thought it did the correct thing. In contrast, an agent error should usually not cause the agent to lecture the user about how they should do the task.

Explanation Causes vs. Triggers

It is important to understand the underlying causes of the need for an explanation to know what to include in an explanation and when to give it. However, to know when to explain, the agent needs to be aware that one of the causes holds true. Here, we differentiate between the underlying causes that make an explanation necessary and the triggers that, in practice, lead to the agent providing an explanation. These triggers include:

- the agent predicting or detecting one of the causes to be true, for example, predicting the user's model to be misaligned with its own model of the world (e.g., [9]) or realizing it made an error (e.g., [55]);
- the user asking an explanation-demanding question;
- the agent having one of the social intentions such as to convince or deceive (e.g., [65]);
- the agent detecting the user's confusion and need for an explanation based on non-verbal cues (e.g., [55]).

As with explanation causes before, there is no strict mapping from triggers to types, but various combinations are possible.

7 Discussion

Firstly, we have shown the types of explanations that are important during human–agent interactions. On the top level of our taxonomy, we distinguished between assistive and introspective explanations. Both types were further analyzed and divided into sub-types. Each of the identified sub-types can be further broken down using the existing classification schemes presented in the related work section. The presented classification schemes can help roboticists design or deploy appropriate explanation capabilities. Firstly, the taxonomy offers a starting point for identifying robot functionality that might benefit from being explainable. Secondly, once candidate modules or contexts have been identified, the annotated collection of papers provides a set of current explainability solutions ready to be used or to be built upon. The suggested explanation types are of different importance depending on the application. For instance, conversational explanations might be more applicable to robot-assisted learning and robots with complex speech capabilities, as in the case when they are connected to a large language model, for instance, a GPT-variant [69]. Assistive explanations become important when the user is not yet familiar with what to do, the task is of high complexity, or the task introduces novel elements regularly. Considering future robots that have a multitude of abilities and take care of many different tasks, it is desirable that they can give a variety of explanations covering multiple types specified in the taxonomy.

Secondly, we have shown why the need for an explanation can arise during human–agent interactions. We differentiate between a variety of agent-centered, user-centered, and external causes that make an explanation become necessary. Combined with the explanation type taxonomy, this classification aids the design of agents that give explanations in the correct situations, exactly when needed, targeting the right underlying cause of the need for explanation. Knowing when explanations become relevant is crucial when making an interactive, collaborative robot explainable, as it implies that the robotic system needs to be able to detect or track these causes of the need for an explanation. For instance, a decision-making system needs to have an uncertainty measure in

order to be able to start an explanatory dialogue once the uncertainty falls under a certain threshold. Similarly, to explain after an error, a robot's modules need to report back failures, or the robot needs to be able to understand that it did something wrong from other input sources like its visual understanding or the collaborator's reaction. These are not trivial capabilities but deeply ingrained in the robot's overall design and architecture.

Assistive explanations

Assistive explanations are rarely discussed in XAI literature and often take the form of how- and what- instead of why-questions. Thus, one might ask whether these are truly explanations or just the simple act of providing a fact, as in providing the current time. However, assistive explanations can provide insights targeting different explanatory levels as they can provide: (1) factual accounts of what has to be done, (2) mechanistic accounts of how the human can help achieve this goal, and (3) functional accounts of why the goals have to be achieved. Thus, assistive explanations should be considered when looking at explanations in the context of human–robot collaborations.

Confusion and Curiosity

We discussed the mental state of feeling the need for an explanation and what causes it. Others discussed giving explanations in the context of the user being confused [55] or curious [27]. Thus, the question arises of how the need for an explanation is connected to the phenomena of confusion and curiosity. When analyzing the predictors of the need for an explanation, Liquin and Lombrozo show that they are related but still distinct to predictors for curiosity [27]. This makes sense as one can feel curious about something that does not need an explanation, for instance, someone's age. Moreover, someone can feel the need for an explanation without being curious. For example, when someone requires an explanation about whether a system adheres to the law, one does not necessarily feel curiosity. The same holds for confusion. A person can feel confused when they do not understand a specific word someone said, without needing an explanation of what was said but just requiring the exact word. Often, however, those feelings go hand in hand. Another major difference is that the need for an explanation can also be external to the user, whereas confusion and curiosity are always mental states of the user. For example, if the agent realizes it did something wrong or is uncertain how to do something, it can give an explanation to accrue potential benefits such as better team performance and increased trust, even if the user did not feel the need for an explanation in that scenario.

8 Conclusion

In this paper, we presented a classification of situations in human–agent interactions in which the agent should be able to provide explanations. Moreover, we looked at what it is that causes the need for an explanation during such interactions in the first place. We validated and refined our findings via a PRISMA scoping review of the field. Our survey can guide robotics practitioners and researchers who want to create agents that can provide explanations as it firmly grounds and contextualizes the concept of explanations within human–agent interaction research. Building on this work, we started to validate when people feel the need for an explanation with actual users who are confronted with human–robot interaction scenarios [70]. In the future, we plan to propose explainable agent architectures taking our findings into account.

Acknowledgments. This work was supported by UKRI (EP/S023356/1) CDT in Safe and Trusted AI, COHERENT (EP/V062506/1), LISI (EP/V010875/1), and the King's Institute for AI. G. Canal was supported by the Royal Academy of Engineering and the Office of the Chief Science Adviser for National Security under the UK Intelligence Community Postdoctoral Research Fellowship programme.

Declarations

There are no conflicts of interest.

References

[1] Rosenfeld, A., Richardson, A.: Explainability in Human–Agent Systems. AAMAS **33** (2019)

[2] Anjomshoae, S., Najjar, A., Calvaresi, D., Främling, K.: Explainable agents and robots: Results from a systematic literature review. In: AAMAS (2019)

[3] Sakai, T., Nagai, T.: Explainable autonomous robots: a survey and perspective. Advanced Robotics **36** (2022)

[4] Setchi, R., Dehkordi, M.B., Khan, J.S.: Explainable robotics in human-robot interactions. Procedia Computer Science (2020)

[5] Sado, F., Loo, C.K., Liew, W.S., Kerzel, M., Wermter, S.: Explainable Goal-Driven Agents and Robots - A Comprehensive Review. ACM Comput. Surv. (2022)

[6] Papagni, G., Koeszegi, S.: Understandable and trustworthy explainable robots: a sensemaking perspective. Paladyn (2021)

[7] Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access **6** (2018)

[8] Sheridan, T.B.: Humans and Automation: System Design and Research Issues. John Wiley & Sons, Inc., USA (2002)

[9] Soni, U., Sreedharan, S., Kambhampati, S.: Not all users are the same: Providing personalized explanations for sequential decision making problems. In: IROS (2021)

[10] Kopecka, H., Such, J.: Explainable ai for cultural minds. In: Workshop on Dialogue, Expl. and Argumentation for HAI (2020)

[11] Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence **267** (2019)

[12] Marr, D.: Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information, (1982)

[13] Srinivasan, R., Chander, A.: Explanation perspectives from the cognitive sciences—a survey. In: IJCAI (2021)

[14] Malle, B.F.: How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction. MIT press, England (2006)

[15] Malle, B.: A coding scheme for folk explanations of behavior (2014)

[16] Panisson, A.R., Engelmann, D.C., Bordini, R.H.: Engineering explainable agents: an argumentation-based approach. In: Engineering Multi-Agent Systems: 9th International Workshop (2022). Springer

[17] Walton, D., Reed, C., Macagno, F.: Argumentation Schemes. Cambridge University Press, England (2008)

[18] Lombrozo, T.: The structure and function of explanations. Trends in cognitive sciences **10** (2006)

[19] Explanation, Cambridge Dictionary. Cambridge Dictionary. accessed Nov, 2022. dictionary.cambridge.org/dictionary/english/explanation

[20] Explaining, Merriam Webster. accessed Nov, 2022. https://www.merriam-webster.com/dictionary/explaining

[21] Ginet, C.: Reasons explanation: Further defense of a non-causal account. The Journal of Ethics **20** (2016)

[22] Faye, J.: The pragmatic-rhetorical theory of explanation. In: Rethinking Explanation. Springer, Dordrecht (2007)

[23] Ehsan, U., Riedl, M.O.: Social Construction of XAI: Do We Need One Definition to Rule Them All? preprint arXiv:2211.06499 (2022)

[24] Wallkötter, S., Tulli, S., Castellano, G., Paiva, A., Chetouani, M.: Explainable embodied agents through social cues: A review. J. Hum.-Robot Interact. **10**(3) (2021) https://doi.org/10.1145/3457188

[25] Wilkinson, S.: Levels and kinds of explanation: lessons from neuropsychiatry. Frontiers in Psychology (2014)

[26] Keil, F.C.: Explanation and understanding. Annu. Rev. Psych. (2006)

[27] Liquin, E., Lombrozo, T.: Determinants and consequences of the need for explanation. In: CogSci (2018)

[28] Krause, L., Vossen, P.: When to explain: Identifying explanation triggers in human-agent interaction. In: INLT for XAI (2020)

[29] Tricco, A.C., Lillie, E., Zarin, W., O'Brien, K.K., Colquhoun, H., Levac, D., Moher, D., Peters, M.D., Horsley, T., Weeks, L., *et al.*: Prisma extension for scoping reviews (prisma-scr): checklist and explanation. Annals of internal medicine **169**(7), 467–473 (2018)

[30] Chari, S., Seneviratne, O., Gruen, D.M., Foreman, M.A., Das, A.K., McGuinness, D.L.: Explanation ontology: a model of explanations for user-centered ai. In: International Semantic Web Conference (2020)

[31] Wilson, J.R., Aung, P.T., Boucher, I.: Enabling a social robot to process social cues to detect when to help a user. preprint arXiv:2110.11075 (2021)

[32] Raymond, A., Gunes, H., Prorok, A.: Culture-based explainable human-agent deconfliction. AAMAS (2020)

[33] Gao, X., Gong, R., Zhao, Y., Wang, S., Shu, T., Zhu, S.-C.: Joint Mind Modeling for Explanation Generation in Complex Human-Robot Collaborative Tasks. In: IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), pp. 1119–1126 (2020). https://doi.org/10.1109/RO-MAN47096.2020.9223595

[34] Hu, S., Chew, E.: The investigation and novel trinity modeling for museum robots. In: Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality (2020)

[35] Schodde, T., Hoffmann, L., Stange, S., Kopp, S.: Adapt, explain, engage—a study on how social robots can scaffold second-language learning of children. ACM THRI (2019)

[36] Gunning, D., Aha, D.: DARPA's Explainable Artificial Intelligence (XAI) Program. AI Magazine (2019)

[37] Fox, M., Long, D., Magazzeni, D.: Explainable planning. In: IJCAI Workshop on Explainable Planning (2017)

[38] Puiutta, E., Veith, E.M.: Explainable reinforcement learning: A survey. In: ML and Knowledge Extraction (2020). Springer

[39] Brandao, M., Mansouri, M., Mohammed, A., Luff, P., Coles, A.: Explainability in multi-agent path/motion planning: User-study-driven taxonomy and requirements. In: AAMAS (2022)

[40] Brandão, M., Canal, G., Krivić, S., Magazzeni, D.: Towards providing explanations for robot motion planning. In: ICRA (2021). https://doi.org/10.1109/ICRA48506.2021.9562003

[41] Buhrmester, V., Münch, D., Arens, M.: Analysis of explainers of black box deep neural networks for computer vision: A survey. ML and Knowledge Extraction (2021)

[42] Tan, H.: Fractual projection forest: Fast and explainable point cloud classifier. In: Winter Conf. on Applications of Computer Vision (2023)

[43] Gao, R., Tian, T., Lin, Z., Wu, Y.: On explainability and sensor-adaptability of a robot tactile texture representation using a two-stage recurrent networks. In: IROS (2021). IEEE

[44] Antonucci, A., Papini, G.P.R., Bevilacqua, P., Palopoli, L., Fontanelli, D.: Efficient prediction of human motion for real-time robotics applications with physics-inspired neural networks. IEEE Access (2021)

[45] Vice, J., Khan, M.M.: Toward accountable and explainable artificial intelligence part two: The framework implementation. IEEE Access (2022)

[46] Bharadhwaj, H.: Layer-wise relevance propagation for explainable deep learning based speech recognition. In: ISSPIT (2018)

[47] Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., Sen, P.: A survey of the state of explainable ai for natural language processing. In: AACL-IJCNLP (2020)

[48] Mota, T., Sridharan, M., Leonardis, A.: Integrated commonsense reasoning and deep learning for transparent decision making in robotics. SN Computer Science **2**(4), 242 (2021)

[49] Coruhlu, G., Erdem, E., Patoglu, V.: Explainable robotic plan execution monitoring under partial observability. IEEE Transactions on Robotics **38**(4), 2495–2515 (2021)

[50] Alvanpour, A., Das, S.K., Robinson, C.K., Nasraoui, O., Popa, D.: Robot failure mode prediction with explainable machine learning. In: CASE (2020). IEEE

[51] Kaptein, F., Broekens, J., Hindriks, K., Neerincx, M.: Evaluating cognitive and affective intelligent agent explanations in a long-term health-support application for children with type 1 diabetes. In: 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 1–7 (2019). IEEE

[52] Abdulrahman, A., Richards, D.: Modelling therapeutic alliance using a user-aware explainable embodied conversational agent to promote treatment adherence. In: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, pp. 248–251 (2019)

[53] Pal, P., Clark, G., Williams, T.: Givenness hierarchy theoretic referential choice in situated contexts. In: Proceedings of the Annual Meeting of the Cognitive Science Society (2021)

[54] Kontogiorgos, D., van Waveren, S., Wallberg, O., Pereira, A., Leite, I., Gustafson, J.: Embodiment effects in interactions with failing robots. In: Conf. on Human Factors in Computing Systems (2020)

[55] Wachowiak, L., Tisnikar, P., Canal, G., Coles, A., Leonetti, M., Celiktutan, O.: Analysing eye gaze patterns during confusion and errors in human–agent collaborations. In: RO-MAN (2022). IEEE

[56] Mirnig, N., Stollnberger, G., Miksch, M., Stadler, S., Giuliani, M., Tscheligi, M.: To err is robot: How humans assess and act toward an erroneous social robot. Frontiers in Robotics and AI (2017)

[57] Kim, T., Hinds, P.: Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. In: RO-MAN (2006)

[58] Das, D., Banerjee, S., Chernova, S.: Explainable ai for robot failures: Generating explanations that improve user assistance in fault recovery. In: HRI (2021)

[59] Sreedharan, S., Srivastava, S., Smith, D., Kambhampati, S.: Why can't you do that HAL? Explaining unsolvability of planning tasks. In: IJCAI (2019)

[60] Han, Z., Phillips, E., Yanco, H.A.: The Need for Verbal Robot Explanations and How People Would Like a Robot to Explain Itself. Transactions on Human-Robot Interaction (2021)

[61] Molineaux, M., Klenk, M., Aha, D.: Goal-driven autonomy in a navy strategy simulation. In: 24th Conference on AI (2010)

[62] Mirnig, N., Giuliani, M., Stollnberger, G., Stadler, S., Buchner, R., Tscheligi, M.: Impact of Robot Actions on Social Signals and Reaction Times in HRI Error Situations. In: Social Robotics, (2015)

[63] Schank, R., Abelson, R.: Scripts, Plans, Goals, and Understanding: An Inquiry Into Human Knowledge Structures. Psychology Press, USA (2013)

[64] Pelikan, H., Hofstetter, E.: Managing delays

in human-robot interaction. ACM Transactions on Computer-Human Interaction (2022)

[65] Rosenfeld, A., Kraus, S.: Strategical argumentative agent for human persuasion. In: 22nd European Conf. on Artificial Intelligence, (2016)

[66] Sreedharan, S., Chakraborti, T., Kambhampati, S.: Foundations of explanations as model reconciliation. Artificial Intelligence (2021)

[67] Väänänen, K., Pohjola, H., Ahtinen, H.-L.A.: Exploring the user experience of artificial intelligence applications: User survey and human-ai relationship model. In: CHI'19 Workshop Where Is the Human? Bridging the Gap Between AI and HCI (2019)

[68] Roque, A., Damodaran, S.K.: Explainable ai for security of human-interactive robots. International Journal of Human–Computer Interaction, 1789–1807 (2022)

[69] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. NeurIPS (2020)

[70] Wachowiak, L., Fenn, A., Kamran, H., Coles, A., Celiktutan, O., Canal, G.: When do people want an explanation from a robot? In: Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction. HRI '24, pp. 752–761. Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3610977.3634990