Explainable Embodied Agents Through Social Cues: A Review

Sebastian Wallkötter, Silvia Tulli, Ginevra Castellano, Ana Paiva, Mohamed Chetouani

▶ To cite this version:

Sebastian Wallkötter, Silvia Tulli, Ginevra Castellano, Ana Paiva, Mohamed Chetouani. Explainable Embodied Agents Through Social Cues: A Review. ACM Transactions on Human-Robot Interaction, 2021, 10 (3), 10.1145/3457188 . hal-03152010

HAL Id: hal-03152010

https://hal.sorbonne-universite.fr/hal-03152010

Submitted on 25 Feb 2021

Explainable Embodied Agents Through Social Cues: A Review

SEBASTIAN WALLKÖTTER[*] and SILVIA TULLI[*], Uppsala University and INESC-ID - Instituto Superior Técnico
GINEVRA CASTELLANO, Uppsala University, Sweden
ANA PAIVA, INESC-ID and Instituto Superior Técnico, Portugal
MOHAMED CHETOUANI, Institute for Intelligent Systems and Robotics, CNRS UMR 7222, Sorbonne Université, France

The issue of how to make embodied agents explainable has experienced a surge of interest over the last three years, and, there are many terms that refer to this concept, e.g., transparency or legibility. One reason for this high variance in terminology is the unique array of social cues that embodied agents can access in contrast to that accessed by non-embodied agents. Another reason is that different authors use these terms in different ways. Hence, we review the existing literature on explainability and organize it by (1) providing an overview of existing definitions, (2) showing how explainability is implemented and how it exploits different social cues, and (3) showing how the impact of explainability is measured. Additionally, we present a list of open questions and challenges that highlight areas that require further investigation by the community. This provides the interested reader with an overview of the current state-of-the-art.

CCS Concepts: • **Computing methodologies** → **Intelligent agents**; *Reasoning about belief and knowledge*; • **Computer systems organization** → **Robotics**; • **General and reference** → *Metrics*; • **Human-centered computing** → *User models.*

Additional Key Words and Phrases: transparency, interpretability, explainability, accountability, expressive behaviour, intelligibility, legibility, predictability, explainable agency, embodied social agents, robots

ACM Reference Format:
Sebastian Wallkötter, Silvia Tulli, Ginevra Castellano, Ana Paiva, and Mohamed Chetouani. 2019. Explainable Embodied Agents Through Social Cues: A Review. 1, 1 (February 2019), 24 pages.

1 INTRODUCTION

Embodied agents are capable of engaging in face-to-face interaction with humans through both, verbal and non-verbal behaviour. They are employed in situations in which joint activities occur, requiring teammates to be able to perceive, interpret, and reason about intentions, beliefs, desires, and goals to perform the right actions. However, even if an embodied agent is endowed with communicative behaviours, merely having them does not guarantee that its actions are understood correctly.

[*]Both authors contributed equally to this research.

Authors' addresses: Sebastian Wallkötter, sebastian.wallkotter@it.uu.se; Silvia Tulli, silvia.tulli@gaips.inesc-id.pt, Uppsala University and INESC-ID - Instituto Superior Técnico, Ginevra Castellano, Uppsala University, Sweden; Ana Paiva, INESC-ID and Instituto Superior Técnico, Lisbon, Portugal; Mohamed Chetouani, Institute for Intelligent Systems and Robotics, CNRS UMR 7222, Sorbonne Université, Paris, France, .

Enabling an embodied agent to use its own behaviours to be better understood by human partners is studied as explainability (among other terms) and is an active research area (see figure 1). Embodied agents have access to differing social cues compared to their non-embodied counterparts, which sets them apart from artificial intelligence (AI) systems that are not embodied. These social cues are handy, because users - particularly non-experts - often have limited understanding of the underlying mechanisms by which an embodied agents choose their actions. Revealing such mechanisms builds trust and is considered ethical in AI [35] and robotics alike. Embodiment-related social cues may help in this context to efficiently communicate such underlying mechanisms for a better interaction experience. Consequently, it is interesting to investigate explainability with a special focus on embodiment.

In human-robot interaction (HRI), embodiment often refers to the use of physical embodiment; however, testing a single aspect of the interaction in online studies is also common, such as the observation of the agent behavior in a simulation or the intepretation of its textual explanations. While this might introduce a confounding difference between the settings (e.g., not accounting for the effect of the embodiment) [50], in such studies, the risk of introducing a confounding factor is outweighed by the possibility isolating the aspect of interest. In explainability research studies with virtual agents manipulating a single behaviour happens frequently, and typically with the aim of adding the investigated aspect to a physically embodied agent later during the research. Hence, we chose to not only review physically embodied agents but also include virtual studies.

Although the concept of explainable embodied agents has become increasingly prevalent, the idea of explaining a system's action is not new. Starting in the late 1970s scholars already began to investigate how expert systems [66, 74, 78] or semantic nets [29, 79], which use classical search methods, encode their explanations in human readable form. Two prominent examples of such systems are MYCIN [74], a knowledge-based consultation program for the diagnosis of infectious diseases, and PROLOG, a logic-based programming language. MYCIN is particularly interesting because it has already explored the idea of interactive design to allow both, inquiry about the decision in the form of specific questions, and rule-extraction based on previous decision criteria [74].

Connecting to this history, explainable embodied agents can be considered a subcategory of explainable AI systems, as they use techniques similar to those mentioned above but with the aim of interacting autonomously with humans and the environment. Therefore, explainable embodied agents should be able to explain their reasoning (justification), explain their view of the environment (internal state) or explain their plans (intent). These abilities can, for example, support the collaboration between agents and humans or improve the agent's learning by aiding the human teacher in selecting informative instruction [17]. Furthermore, explainable AI techniques might exploit attention direction to communicate points of confusion for embodied agents using gaze and verbalization [60]. Some examples in the literature have proposed generating explanations of reasoning processes using natural language templates Wang et al. [75]. In contrast, other work focuses on making the actions of the embodied agents explicit by design (i.e., legible) [22].

In this paper, we want to show how these unique social cues can be used for building explainable embodied agents, and highlight which aspects require further inquiry. As such, **we contribute** to the field of explainability with **a review of the existing literature on explainability** that organizes it by (1) providing an overview of existing definitions, (2) showing how explainability is implemented and how it exploits different social cues, and by (3) showing how the effect of explainability is measured. This review aim to provide interested readers with an overview of the current state-of-the-art. Additionally, we present **a list of open questions and challenges** that highlight areas that require further investigation by the community.

1.1 Similar Reviews

Other authors have previously written about explainability in the form of position papers and reviews [24, 26, 38, 53, 68, 71].

Doshi-Velez and Kim [21] and Lipton [51] sought to refine the discourse on interpretability by identifying the desiderata and methods of interpretability research. Their research focused on the interpretation of machine learning systems from a human perspective and identified trust, causality, transferability, informativeness, and fair and ethical decision-making as key aspects of their investigation. Rosenfeld and Richardson [63] provided a notation for defining explainability in relation to related terms such as interpretability, transparency, explicitness, and faithfulness. Their taxonomy encompasses the motivation behind the need for explainability (unhelpful, beneficial, critical), the importance of identifying the target of the explanation (regular user, expert user, external entity), when to provide the explanation, how to measure the effectiveness of the explanation and which interpretation of the algorithm has been used. Anjomshoae et al. [4] conducted a systematic review on the topic of explainable robots and agents and clustered the literature with respect to the user's demographics, the application scenario, the intended purpose of an explanation, and whether the study was grounded in social science or had a psychological background. The review summarized the methods used to implement the explanation to the user with its dynamics (context-aware, user-aware, both or none), and the types of explanation modality. Alonso and De La Puente [3] proposed a review of the system's explainability in a shared autonomy framework, stressing the role of explainability in flexible and efficient human-robot collaborations. Their review underlines how explainability should vary in relation to the level of system autonomy and how the exploitation of explainability mechanisms can result from an explanation, a property of an interface, or a mechanical feature.

Other reviews explored the cognitive aspects of explainability by targeting the understanding of behaviour explanations and how this helps people find meaning in social interaction with artificial agents. The book by Malle [54] argued that people expect explanations using the same conceptual framework used to explain human behaviours. Similarly, the research of Miller [55] focused on the definition of explanations in other relevant fields, such as philosophy, cognitive psychology/science, and social psychology. The author described an explanation as contrastive, selected and social, specifying that the most likely explanation is not always the best explanation for a person.

1.2 Methodology

For this review, we chose to use a keyword based search in Scopus[1] database to identify relevant literature, as this method makes our search reproducible.

First, we identified a set of relevant papers in an unstructured manner based on previous knowledge of the area. From each paper, we extracted both, the indexed and the author keywords, and rank ordered each keyword by occurrence. Using this method, we identified key search terms such as *human-robot interaction, transparent, interpretable, explainable,* or *planning*.

We then grouped these keywords by topic (see table 1) and performed a pilot search on each topic to determine how many of the initially identified papers were recovered. We then combined each group using *AND*, which led to a corpus of 263 papers[2]. All authors participated in this initial extraction process.

[1] https://www.scopus.com/

[2] A spreadsheet detailing the raw search results can be found in the supplementary files.

Table 1. Inclusion Criteria and Search String

Topic	Description	Search Term
Human Involvement	Exclude papers without human involvement, e.g., position papers or agent-agent interaction	("human-robot" OR "child-robot" OR "human-machine")
Explainability		(transparen* OR interpretabl* OR explainabl*)
Explainability II		(obser* OR legib* OR visualiz* OR (commun* AND "non-verbal"))
Autonomy	Exclude papers that are not using an autonomous agent	(learn* OR plan* OR reason* OR navigat* OR adapt* OR personalis* OR decision-making OR autonomous)
Social Cues	Exclude papers that do not have a social interaction between human and agent	(social OR interact* OR collab* OR shared OR teamwork OR (model* AND (mental OR mutual)))
Agent	Exclude papers that do not use an agent	(agent* OR robot* OR machine* OR system*)
Recency	Only consider the last 10 years	(LIMIT-TO (PUBYEAR, 2019) OR ... OR LIMIT-TO (PUBYEAR, 2009))
Subject Area	Only consider papers from computer science, engineering, math, psychology, social sciences, or neuroscience	(LIMIT-TO (SUBJAREA, "COMP") OR LIMIT-TO (SUBJAREA, "ENGI") OR LIMIT-TO (SUBJAREA, "MATH") OR LIMIT-TO (SUBJAREA, "SOCI") OR LIMIT-TO (SUBJAREA, "PSYC") OR LIMIT-TO (SUBJAREA, "NEUR"))
Full Search String	TITLE-ABS-KEY ("human-robot" OR "child-robot" OR "human-machine") AND (transparen* OR interpretabl* OR explainabl*) AND (obser* OR legib* OR visualiz* OR (commun* AND "non-verbal")) AND (learn* OR plan* OR reason* OR navigat* OR adapt* OR personalis* OR decision-making OR autonomous) AND (social OR interact* OR collab* OR shared OR teamwork OR (model* AND (mental OR mutual))) AND (agent* OR robot* OR machine* OR system*) AND (LIMIT-TO (PUBYEAR, 2019) OR LIMIT-TO (PUBYEAR, 2018) OR LIMIT-TO (PUBYEAR, 2017) OR LIMIT-TO (PUBYEAR, 2016) OR LIMIT-TO (PUBYEAR, 2015) OR LIMIT-TO (PUBYEAR, 2014) OR LIMIT-TO (PUBYEAR, 2013) OR LIMIT-TO (PUBYEAR, 2012) OR LIMIT-TO (PUBYEAR, 2011) OR LIMIT-TO (PUBYEAR, 2010) OR LIMIT-TO (PUBYEAR, 2009)) AND (LIMIT-TO (SUBJAREA, "COMP") OR LIMIT-TO (SUBJAREA, "ENGI") OR LIMIT-TO (SUBJAREA, "MATH") OR LIMIT-TO (SUBJAREA, "SOCI") OR LIMIT-TO (SUBJAREA, "PSYC") OR LIMIT-TO (SUBJAREA, "NEUR"))	

Next, we manually filtered this list to further remove unrelated work by judging relevance based on titles, abstracts, and full text reads. To ensure selection reliability, both main authors rated inclusion of each paper independently. If both labelled the paper as relevant, we included the paper; similarly, if both labelled it as unrelated, we excluded it. For papers with differing decisions, we discussed their relevance and made a joint decision regarding the paper's inclusion. This left us with 32 papers for the final review [3].

For the excluded papers, each main author indicated why a paper was excluded for the following reasons:

[3] A spreadsheet detailing the reason for excluding a paper can be found in the supplementary files.

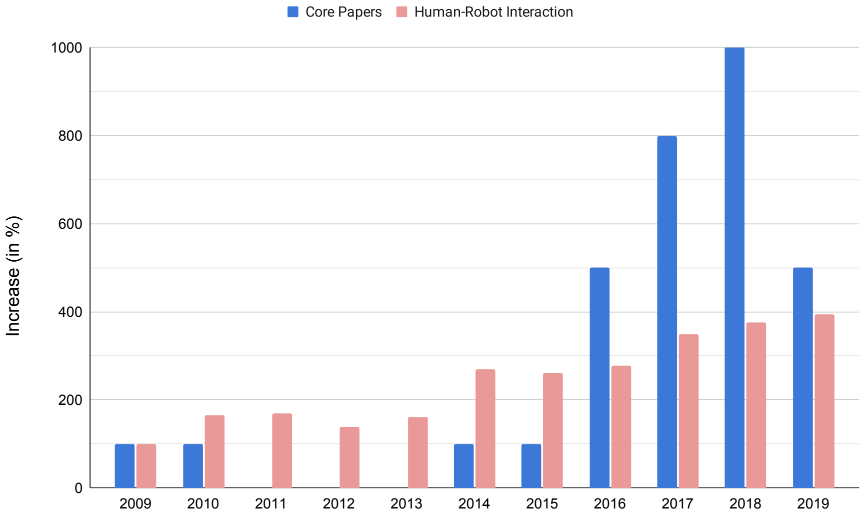

Fig. 1. Percentage increase in the number of publications relative to the first year (2009) in the reviewed area. For comparison, the graph also shows the percentage increase in the number of publications overall measured by the number of publications indexed in Scopus that use the keyword "Human-Robot Interaction".

- **no article** The paper was a book chapter or review paper. (~ 15.63% excluded)
- **wrong topic** The paper presented work in a different focus area, e.g, material science, teleoperation, or generically making robots more expressive (without considering explainability). (~ 45.98% excluded)
- **wrong language** The paper was not written in English. (~ 0.46% excluded)
- **no embodiment** The paper did not consider an embodied agent. (~ 11.26% excluded)
- **no autonomy** The paper did not consider autonomous embodied agents. (~ 13.33% excluded)
- **no social interaction** The paper did not investigate explainability in a context where a human was present. (~ 13.10% excluded)

Figure 1 shows a comparison between the publication rate of the identified papers (core papers) and the publication rate of the general field, which is measured, as a proxy, by papers published using the keyword *human-robot interaction*. We can see that there is a growing interest in the topic reviewed here.

2 DEFINITIONS

2.1 Review of Definitions

Before starting to look into how different social cues are exploited, it is important to understand what different scholars mean when they talk about explainability.

Among the surveyed papers, we analyzed the terms used in the title. We extrapolated the root terms and categorized them (e.g. explanation -> explainability, expressive -> expressivity). Assuming that papers with the same authors share the same definition, we grouped them by author, and only report the most recent definition in Table 6. When the definition was from an outside source not cited before, we mentioned both the definition and the reference.

Some authors that refer to transparency identify transparency as the process or the capability of revealing hidden or unclear information about the robot. Akash et al. [1][2] and Ososky et al. [58] cited the definition of Chen et al. [18], which named transparency the descriptive quality of an interface. The descriptiveness of the interface affects the operator on three levels (perception, comprehension, and projection) leveraged by Endsley's model of situation awareness [23]. Ososky et al. [58] further refer to the dictionary definition of *transparency*[4]; thus, they chose the property of being *able to be seen through* or *easy to notice or understand* as their definition. Chao et al. [17] gave a similar definition in the context of robot active learning referring to transparency as revealing to the teacher unknown information about the robot.

In an extension, Floyd and Aha [27] and Hayes and Shah [34] refer to both explainability and transparency. Hayes and Shah [34] describe explainability as the embodied agent's ability to synthesize policy descriptions and respond to human collaborators.

Starting from the research of Kulkarni et al. [43] and Zhang and Liu [82], Chakraborti et al. [16] and [14] formulated their idea of explainability as a model reconciliation problem. They use explanations to move the human model of the robot to be in conformance with the embodied agent model. Gong and Zhang [30] differentiate their work by shifting the interest in signalling robot intentions before actions occur. Along similar lines, Arnold et al. [6] defined explainability as a policy update given by the robot to the human to reduce the likelihood of costly or dangerous failures during joint task execution.

Baraka and Veloso [8] employed the term *expression* for *externalizing hidden information of an agent*. The work of Kwon et al. [44] extended this notion of expressivity by targeting the communication of robot incapability. To do so, the robot should reveal the intentions and the cause of its failures. The same concept is referred to using the word communicability; e.g., Huang et al. [37] refer to *communicate* for describing the robot capability of expressing its objectives and the robots capability to enable end-users to correctly anticipate its behaviour.

Similarly, Schaefer et al. [64] investigated the *understandability* of the embodied agent's intentions for effective collaborations with humans. Following this idea, Grigore et al. [32] referred to *predictability*, building upon hidden state representation.

Other studies in the literature do not refer to a specific capability of their system in the title but highlighted the application scenarios (e.g., autonomous driving [12], human robot teams in the army [80], interactive robot learning [52]) or mentioned *behavioural dynamics* [45] and *human-machine confidence* [49].

Although there exists large diversity and inconsistency in the language, there seem to be commonalities in what the authors identify as explainability. Furthermore, other authors that use terms such as transparency, expressivity, understandability, predictability and communicability present definitions that are congruent with those provided for explainability. We noticed that all the given definitions share the following aspects: (1) they all refer to an embodied agent's capability or system's module, (2) they all specify that what should be explained/signalled are the internal workings of the robot (e.g. agent's intent, policy, plan, future plans), and (3) they all consider the human as a target of the explanations.

[4] https://www.merriam-webster.com

Table 2. Papers by Terminology Used in the Title

Category	Paper
Transparency	[1, 2, 10, 17, 18, 26, 27, 60–62] [28, 34]
Explainability	[6, 14, 16, 30, 75–77], [28, 34]
Expressivity	[8, 44, 83]
Understandability	[64, 65]
Predictability	[32, 67]
Communicability	[36, 48]
None	[12, 39, 45, 49, 52]

2.2 Our Definition

Therefore, we provide a definition that aims to be comprehensive for our literature. **We define the explainability of embodied social agents as their ability to provide information about their inner workings using social cues, such that an observer (target user) can infer how/why the embodied agent behaves the way it does.**

2.3 Motivation

While investigating the definitions, we noticed that the motivation of the experiment plays a key role in the choice of a specific definition. In particular, we identified the following reasons for investigating explainability:

- **Interactive machine/robot learning** investigates the need of explainability in the context of robot learning. The main idea is that revealing the embodied agent's internal states allows the human teacher to provide more informative examples [6, 17, 52, 73].
- **Human Trust** states that adding explainability increases human trust and system reliability. This motivation empathizes the importance of communicating the agent's uncertainty, incapability, and existence of internally conflicting goals [44, 62, 64, 76].
- **Teamwork** underlines the value of explainability in human-robot collaboration scenarios to build shared mental models and predict the embodied agent's behaviour [16, 34, 36, 49, 65].
- **Ethical decision-making** suggests that communicating the embodied agent's decision-making processes and capabilities, paired with situational awareness, increases a user's ability to make good decisions [1, 44, 61].

We have aggregated the individual definitions used in each paper and the motivations behind them in table 6. Guidance and dialogue with a human tutor are aspects that are important for interactive machine/robot learning ("dialog to guide human actions" [52], "revealing to the teacher what is known and what is unclear" [17]). Providing information about the level of uncertainty and expressing robot incapability are core concepts of explainability that enhance human trust ("communicate uncertainty" [76], "provide appropriate measures of uncertainty" [62], "express robot incapability" [44]). The ability to anticipate an embodied agent's behaviour and establish a two-way collaborative dialogue by identifying relevant differences between the humans' and the robots' model are shared elements of the definitions

Table 3. Papers on Explainability Ordered by Social Cues

Category	Paper
Speech	[6, 26, 52, 60, 62]
Text	[1, 2, 14, 18, 27, 28, 30, 34, 61, 75–77]
Movement	[12, 14, 17, 36, 44, 45, 52, 60, 65, 83]
Imagery	[10, 12, 49, 60, 62, 64]
Other/Unspecified	[8, 60]/ [16, 32, 39, 48]*

around teamwork ("anticipate robot's behaviour" [36], "establish two-way collaborative dialogue" [49], "reconcile the relevant differences between the humans' and robot's model" [13], "share expectations" [34]). Authors that refer to ethical decision-making identify the communication of intentions and context-dependent recommendations as crucial information ("robot's real capabilities, intentions, goals and limitations" [61], "context-dependent recommendations based on the level of trust and workload" [2]).

3 SOCIAL CUES

We have claimed above that embodied agents can become explainable using unique types of social cues that are not available to agents lacking such embodiment. One example is the ability to point to important objects in a scene - assuming the agent has an extremity that affords pointing. A non-embodied agent has to use a different way to communicate the importance of that object.

Hence, we screened the core papers to check which modality the authors deployed to make the agent more explainable. We then logically grouped the core papers based on these types of social cues and identified five groups:

- **Speech** A lexical statement uttered verbally using a text-to-speech mechanism.
- **Text** A lexical statement displayed as a string presented as an element of a typically screen-based user interface.
- **Movement** A movement that is either purely communicative, or that alters an existing movement to make it more communicative.
- **Imagery** A drawing or image (often annotated) presented as an element of a user interface (typically screen-based).
- **Other/Unspecified** All papers that use social cues that do not fit within above set of categories, or where the authors did not explicitly specify the modality (the latter is marked with an asterisk*).

This grouping is shown in table 3. Surprisingly, our search did not yield any papers that investigate non-lexical utterances (beeping noise, prosody, etc.), which was contrary to our expectations. A possible explanation for this could be that our search terms did not capture a broad enough scope, because experiments investigating such utterances may use yet again a different terminology. Another possibility could be that it seems much harder to communicate an explanation through *beeps and boops* instead of using speech; especially when considering the wide availability of text-to-speech synthesizers (TSSs).

The wide availability of TTS synthesizers may also explain another interesting result of this analysis. Many works focus on lexical utterances (speech and text). Potentially, such utterances are seen as easier to work with when giving explanations because of the high expressivity of natural language.

On the other hand, lexical utterances may add additional complexity to the interaction, because a sentence has to be interpreted and understood, whereas, other social cues may be faster/easier to interpret. While there does exist work that investigates the added cognitive load of having explainability versus not having explainability [76], comparing lexical utterances with other forms of explainability is currently underexplored. This prompts the question of whether lexical utterances are always superior to achieve explainability and, if not, under what circumstances other social cues perform better.

4 EXPLAINABILITY MECHANISMS

Next, we report and discuss the methods employed to achieve explainable behaviours in social agents with a specific focus on embodied agents. From an HRI perspective, introducing explainability mechanisms is challenging, as uncertainty is inherent to the whole process of interaction from perception to decision and action. In addition, the methods used to implement explainability require explicit consideration of the human capability to correctly infer the agent goals, intentions or actions from the observable cues.

Looking at one specific implementation, Thomaz and Breazeal [72] introduced the socially guided machine learning (SG-ML) framework which seeks to augment traditional machine learning models by enabling them to interact with humans. Two interrelated questions are investigated: (1) how do people want to teach embodied agents and (2) how do people design embodied agents that learn effectively from natural interaction and instruction. This framework considers a reciprocal and tightly coupled interaction; the machine learner and human instructor cooperate to simplify the task for each other. SG-ML considers explainability to be a communicative act that helps humans understand the machine's internal state, intent, or objective during the learning process.

Interactive situations. As humans interact with robots, explainability becomes a key element. Different interactive scenarios have been explored in the analysed papers, in particular, scenarios where humans shape the behaviour of a robot by providing instructions and/or social cues through interactive learning techniques. Despite of a strong emphasis on interactive robot learning scenarios, there are also other types of tasks as illustrated by Figure 2. For example, the behaviour shaping explored in [40, 57] aims to exploit instructions and/or social cues to steer robot actions towards desired behaviours. Various interaction schemes have been proposed including instructions [33, 59], advice [31], demonstrations [5], guidance [56, 69], and evaluative feedback [41, 56]. Then, computational models, mostly based on machine learning, are exploited to modify agent states s and actions a to achieve a certain goal g.

As mentioned by Broekens and Chetouani [11], most of computational approaches for social agents consider a primary task, e.g., learning to pick an object, and explainability arises as a secondary task by either communicating the agent's internal states, intentions, or future goals. Existing works distinguish the nature of actions performed by the agent, such as task - oriented actions a_T and communication - oriented actions a_C. a_T are used to achieve a goal g such as sorting objects. a_C are used by the agent to communicate with humans such as queries or pointing to objects. This follows from the speech act theory [42], which treats communication as actions that have an intent and an effect (a change of mind by the receiver of the communication).

In such a context, explainability mechanisms are employed to reduce uncertainty during the shaping process using communicative actions a_C before, during, or after performing a task action a_T, which will change the agent's next state s'. The challenge for explainability mechanisms is then to transform agent states s and task oriented actions a_T into communicative actions using either using natural language or non-verbal cues. To tackle this challenge, several explainability mechanisms have been proposed for embodied agents.

Computational paradigms. Various computational paradigms are employed ranging from supervised learning to reinforcement learning. In supervised learning, a machine is trained using data, e.g., different kinds of objects, which are labelled by a human supervisor. In the case of interactive robot learning, the supervisor is a human teacher and provides labelled examples based on embodied agent queries or explanations. In addition, the level of expertise of the human is rarely questioned and considered ground truth. To tackle these challenges, Chao et al. [17] proposed an active learning framework for teaching embodied agents to classify pairs of objects (a_T). Active learning is a type of interactive machine learning that allows the learner to interactively query the supervisor to obtain labels for new data (a_C). By doing so, the robot improves both learning and the explainability by communicating about uncertainty. Often, active learning is a form of semi-supervised learning, that combines human supervision and processing of unlabelled data.

Another paradigm is reinforcement learning (RL) [70], which is one of the three basic machine learning paradigms. Here, an agent acts in an environment, observing its state s and receiving a reward r. Learning is performed by a trial-and-error process through interaction with the environment and leverages the Markov decision process (MDP) framework. MDPs are used to model the agent's policy and help with decision making under uncertainty. This paradigm allows one to represent, plan, or learn an optimal policy - a mapping from current state to action. Analysing this policy provides insights into future and current states and actions. Hayes and Shah [34] developed a policy explanation framework based on the analysis of execution traces of an RL agent. The method generates explanations (a_C) about the learned policy (a_T) in a way that is understandable to humans. In RL, theoretical links between (task) learning schemes and emotional theories could be performed. Broekens and Chetouani [11] investigated how temporal difference learning [70] could be employed to develop an emotionally expressive learning robot that is capable of generating explainable behaviours via emotions.

To increase the understanding of robot intentions by humans, the notion of legibility is often introduced in robotics. Legibility and explainability are considered similar notions that aim to reduce ambiguity over possible goals that might be achieved. One key concept for achieving legibility/explainability is to explicitly consider a model of the human observer, and find plans that disambiguate possible goals. Dragan et al. [22] proposed a mathematical model able to distinguish between legibility and predictability. Legibility is defined as the ability to anticipate the goal, and predictability, is defined as the ability to predict the trajectory. The mathematical model exploits observer expectations to generate legible/explainable plans. Huang et al. [36] propose modelling how people infer objectives from observed behaviour, and then selecting those behaviours that are maximally informative. Inverse reinforcement learning is used to model observer capability of inferring intentions from the observation of agent behaviours. Explainability implementations based on these methods consider that task-oriented actions (a_T) and communicative actions (a_C) are performed through the same channel, e.g., movement of the robot's arm both achieves a task and communicates the goal [67], [65].

Explainability mechanisms. In artificial intelligence (AI), explainability addresses the understanding of the mechanisms by which a model works aiming to reduce the model's *black box* [9]. Deep learning is a typical black-box machine learning method that achieves data representation learning using multiple nonlinear transformations. In contrast, a linear model is considered as explainable since the model is fully understandable and explorable by means of mathematical analysis and methods. In Barredo Arrieta et al. [9], the authors argue that a model is considered to be explainable if by itself it is understandable and propose various levels of model explainability: (1) simulatability, the ability to be simulated or conceptualized strictly by a human, (2) decomposability, the ability to explain each part of the model, and (3) algorithmic transparency, the ability of the user to understand the process followed by the model to produce any given output from its input data.

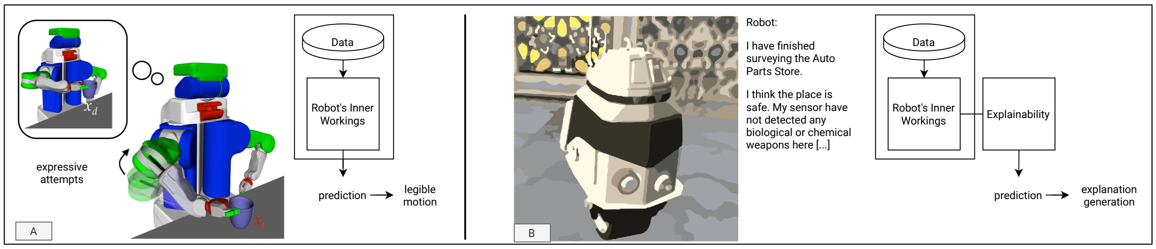

Fig. 2. Example A - Kwon et al. [44], Example B - inspired by Wang et al. [75]. Agent and Explainability Mechanisms. Intrinsic explainability refers to models that are explainable by design. Post-hoc (external) explainability refers to explainability mechanisms that are applied after the decision or execution of actions. Explainability could be either performed by external mechanisms that are separable from the task execution (visualization) or intrinsically computed by the agent policy (e.g. query learning, communicative gestures). Permission to reuse figure A was kindly provided by the authors.

Intrinsic explainability refers to models that are explainable by design (Figure 2). Post-hoc (external) explainability refers to explainability mechanisms that are applied after the decision or execution of actions. Post-hoc methods are decoupled from the model and aim to enhance the explainability of models that are not explainable by design (intrinsic) [51]. Post hoc explainability mechanisms such as visualization, mapping the policy to natural language, or explanation are used to convert a non-explainable model into a more explainable one.

A large body of work aiming to achieve explainability in human-agent interaction does not explicitly refer to definitions that originate from machine learning. Explainability can be either performed by external mechanisms that are separable from the task execution (visualization) [60, 65, 83] or intrinsically computed by the agent policy (e.g. query learning, communicative gestures) [17, 67].

The implementation of explainability can also be done at several levels. For example, via the situation-awareness-based agent transparency (SAT) model, which is based on a Belief, Desire, Intention (BDI) architecture, considers three levels of explainability: Level 1–Basic Information (current status/plan); Level 2–Reasoning Information; Level 3–Outcome Projections [10].

Mapping agent policy ($a_{\mathbf{T}}$) to natural language ($a_{\mathbf{C}}$) is a methodology that is increasingly employed in AI to design explainable AI [9]. In HRI, a similar trend is observed [2, 6, 34, 76]. The challenge will be to map agent policy to both verbal and non-verbal cues (see also section 3).

5 EVALUATION METHODS

Existing works assess the effects of explainability on a variety of measures including but not limited to, self-reported understanding of the agent [30], number of successful task completions [76], number of false decisions [76], task completion time [17], number of irredeemable mistakes [6] or trust in automation [10]. During our review three major categories of measurements emerged:

- **Trust** measures how willing a user is to agree with a decision of a robot - based on the provided explanation -, how confident a user is about the embodied agent's internal workings (internal state), or if the user agrees with the plan provided by the robot (intent). It is measured using a self-report scale.
- **Robustness** measures the avoidance of failure during the interaction. Typically researchers want to determine whether the embodied agent's intent has been communicated correctly. It is often measured observationally, e.g., by counting the frequency of successful achievements of a goal.

Table 4. Papers on Explainability by Measure

Type	Outcome	Papers
Robustness	Positive	[7, 18, 32, 36, 44, 45, 60, 65, 76, 77]
Robustness	Negative	
Robustness	Non-significant	[17]
Robustness	No statistical test	[14, 49, 67]
Trust	Positive	[7, 10, 17, 18, 47, 64, 75, 77, 83]
Trust	Negative	
Trust	Non-significant	[2]
Trust	No statistical test	[1, 6, 16, 27, 28, 30, 44, 61]
Efficiency	Positive	[2, 47, 76]
Efficiency	Negative	
Efficiency	Non-significant	[17, 18, 60, 75]
Efficiency	No statistical test	[16, 62]
other	any	[15, 34, 39, 52, 80]

- **Efficiency** measures how quickly the task is completed. The common hypothesis behind using this measure is that the user can adapt better to a more explainable robot, and form a more efficient team. It is commonly measured by wall clock time, or number of steps until the goal.

Among these measures, trust received the most attention. While there is large variance in which scale is used (often scales are self-made), a common element in the studies is the use of self-report questionnaires.

Although the consensus is that the presence of explainability generally increases trust (see Table 4), how effective a particular social cue is in doing so has received much less attention. Comparisons that do exist often fail to find a significant difference between them [10, 76]. Similarly, due to the large range of mechanisms tested - and the even larger array of scenarios -, there is little work on how robust a specific mechanism performs across multiple scenarios. Hence, while some form of explainability seems to be clearly better then none, which specific mechanism to choose for which specific situation remains an open question.

Less studied, but no less important, is the effect of explainability on the robustness of an interaction. Research on the interplay between explainability and robustness uses tasks where mistakes are possible, and measures how often these mistakes occur [10, 60]. The core idea is that participants create better mental models of the robot when it is using explainability mechanisms. Better models will lead to better predictions of the embodied agent's future behaviour, allowing participants to anticipate low performance of the robot, and to avoid mistakes in task execution. However, experimental evidence on this hypothesis is not always congruent, with the majority of studies showing support for the idea, e.g., [60], and other studies finding no significant difference, e.g., [10]. As the majority does find a positive effect, we can conclude that explainability does help improve reliability, although not in all circumstances. A more detailed account of when it does or does not remains a subject for future experimental work.

Finally, efficiency is a metric that some researchers have considered while manipulating explainability. It has been operationalized by comparing wall clock time until task completion across conditions [17], or time until human response [18]. Of the three types of measures, this type has received the least attention, and the findings are quite mixed. Approximately half of the analysed papers find that making embodied agents explainable makes the team more efficient, while the other half find no difference. However, a clear explanation for these conflicting findings remains a topic of future work.

Table 4 shows the core papers grouped by the evaluation methods discussed above and indicates whether the effect of explainability on it was positive, negative, or non-significant. One important note is that many papers introduce a measurement called *accuracy*; however, usage of this term differs between authors. For example, Chao et al. [17] used accuracy to refer to the embodied agent's performance after a teaching interaction; hence it was being a measure of robustness, whereas Baraka and Veloso [8]'s accuracy referred to people's self-rated ability to predict the robot's move correctly, a measure of trust.

In summary, there is enough evidence that explainability offers a clear benefit to virtual embodied agents in building trust, with some support for physical embodied agents. Additionally, there is evidence that explainability can decrease the chance of an unsuccessful interaction (improve robustness). However, papers looking to improve the efficiency of the interaction find mixed results. A possible explanation for this could be that while explainability makes the interaction more robust, the time added for the embodied agents to display and for the human to digest the additional information nullifies the gain in efficiency.

In addition to the above analysis, this section identified the following open questions: (1) Is a particular explainability mechanism best suited for a specific type of embodied agent, a specific type of scenario, or both? (2) What are good objective measures with which we can measure trust in the context of explainability? (3) Why does explainability have a mixed impact on the efficiency of the interaction?

6 DISCUSSION

In the above sections we provided a focused view on four key aspects of the field: (1) definitions used, and the large diversity thereof, (2) which social cues and (3) algorithms are used to link explainability mechanisms to the embodied agent's state or intent, and (4) the measurements to assess explainability mechanisms. What is missing is a discussion of how these aspects relate to each other when looked at from a 10, 000 foot view, and a discussion of the limitations of our work.

It is almost self-explanatory that the scenario chosen to study a certain explainability mechanism depends on the author's research goal. As such, it is unsurprising that we can find a large diversity of tasks, starting from evaluation in pure simulation [34], or discussions of hypothetical scenarios [47, 61] all the way to joint furniture assembly [62].

The most dominant strand of research has its origin in decision making, and mainly views the robot as a support for human decisions [1, 2, 10, 18, 27, 28, 75–77]. In this line of research, explainability is mostly commonly defined via the SAT-model (i.e., Situation Awareness-Based Agent) [19]. One of the key questions is how much a person will trust the embodied agent's suggestions, based on how detailed the given justification for the embodied agent's decision is. While these studies generally test a virtual agent shaped like a robot, the findings here can be easily generalized to the field of human-computer-interaction (HCI), due to their design. Hence, SAT model-based explanations can help foster trust not only in HRI, but also in the domain of expert systems and AI. Hence, this work partially overlaps with the domain of explainable AI (XAI).

The second strand of research sets itself apart by using humans as pure observers [7, 14–16, 30, 36, 44, 65, 67, 83]. Common scenarios focus on communicating the embodied agent's internal state or intent by having humans observe a physical robot [7, 65, 67] or video recordings/simulations of them [7, 44, 83]. Other researchers choose to show maps of plans generated by the robot and explanations thereof [14, 16, 30, 36]; the researchers' aim here is to communicate the robot's intent. In all scenarios, the goal is typically to improve robustness, although other measures have been tested.

Particularly well done here is the work of Baraka et al. [7], who first describe how to enhance a robot with LED lights to display its internal state, use crowd sourcing to generate expressive patterns for the LEDs, and then validate the pattern's utility in both a virtual and a physical user study. This pattern of having participants - typically from Amazon Mechanical Turk (AMT) - generate expressive patterns in a first survey, and then validate them in a follow-up study was also employed by Sheikholeslami et al. [67] in a pick-and-place scenario. We think that this crowdsourcing approach deserves special attention, as it will likely lead to a larger diversity of candidate patterns compared to an individual researcher generating them. Considering the wide availability of online platforms, such as AMT and Polific, this is a tool that future researchers should leverage.

A third strand of research investigates explainability in interaction between a human and a robot [6, 17, 45, 52, 60, 62, 64] or a human and an AI system [49]. Studies in this strand investigate the impact of different explainability mechanisms on various interaction scenarios and whether they are still useful when the human-robot dyad is given a concrete task. This is important, because users can focus their full attention on the explainable behaviour in the observer setting; in interaction scenarios, on the other hand, they have to divide their attention. Research in this strand is more heterogeneous, likely due to the increased design complexity of an interaction scenario. At the same time, the amount of research done, i.e., the number of papers identified, is less than the research done following the observational design above; probably because of the the above mentioned added complexity. Nevertheless, we argue that more work on this strand is needed, as we consider testing explainability mechanisms in an interaction as the gold standard for determining their utility and effectiveness.

Finally, some researchers examined participants' responses to hypothetical scenarios [47, 61]. The procedure in these studies is to first describe a scenario to participants in which a robot uses an explainability mechanism during an interaction with a human. Then, participants are asked to give their opinion about this interaction, which is used to determine the utility of the mechanism. This method can be very useful during the early design stages of an interaction, and can help find potential flaws in the design before spending much time implementing them on a robot. At the same time, it may be a less optimal choice for the final evaluation, especially when compared to the other methods presented above.

7 CHALLENGES IN EXPLAINABILITY RESEARCH

Shifting the focus to how results are reported in research papers on explainability, we would like to address two challenges we faced while aggregating the data for this review.

The first challenge is the large diversity and inconsistency of language used in the field. Transparency, explainability, expressivity, understandability, predictability and communicability are just a few examples of words used to describe explainability mechanisms. Authors frequently introduce their own terminology when addressing the problem of explainability. While this might allow for a very nuanced differentiation between works, it becomes challenging to properly index all the work done, not only because different authors addressing the same idea may use different terminology but also especially because different authors addressing different ideas end up using the same terminology.

Other reviews on the topic have pointed this out as well [13, 63], and it became a challenge in our review, as we cannot ensure completeness of a keyword search based approach. The most likely cause of this is because the field is seeing rapid growth, and precise terminology is still developing.

This work tries to address this first challenge by showing how different terms are used to identify similar concepts and providing a definition that aims to be comprehensive for the surveyed papers.

The second challenge was that many authors only define the explainability mechanism they investigate implicitly. We often had to refer to the concrete experimental design to infer which mechanism was studied. While all the important information is still present in each paper, we think that explicitly stating the explainability mechanism under study can help discourse regarding explainability become much more concrete.

In extension, some authors have implemented explainability mechanisms on robotic systems that are capable of adapting their behaviour or performing some kind of learning. In many cases, these learning algorithms were unique implementations, or variations of standard algorithms, e.g., reinforcement learning, which make them very interesting. How to best incorporate an explainability mechanism into such a framework is still an open question. Unfortunately, we found that the details of the method are often underreported and that we could not extract enough data on what has been done so far. We understand that this aspect is often not the core contribution of a paper and that space is a constraint. Nevertheless, we would like to encourage future contributions to put more emphasis on how explainability mechanisms are integrated into existing learning frameworks. Technical contributions such as this could prove very valuable for defining a standardized approach to achieve explainability using embodied social agents.

8 OPEN QUESTIONS

While performing the review, we identified a set of open questions. For convenience we enumerate them here:

(1) What are good models to predict/track human expectations/beliefs about the embodied agent's goals and actions?
(2) What are efficient learning mechanisms to include the human in the loop when building explainability into embodied agents?
(3) How does the environment and embodiment influence the choice of social cues used for explainability?
(4) What are good objective measures by which we can measure trust in the context of explainability?
(5) Why does explainability not have a strictly positive impact on the efficiency of the interaction?

9 CONCLUSION

Above we conducted a systematic review of the literature on explainable embodied agents. We used keyword based search to identify 32 relevant contributions and provided a detailed analysis of them.

First, we analysed the definitions of explainability used in each piece, highlighting the heterogeneity of existing definitions and stating our definition. In the process, we identified four main motivations that lead researchers to study explainability: (1) interactive robot/machine teaching, (2) human trust, (3) teamwork, and (4) ethical decision making. We then detailed why explainability is important for each, and identified the motivations and definition behind each of the surveyed papers. Second, we looked at social cues used as vehicles to deliver the explainability mechanism. We identified the categories of (1) speech, (2) text, (3) movement, and (4) imagery and described how each provides explainable behaviours. Third, we took stock of the algorithms used to select which part of the interaction should be made explainable. We found that only a small fraction of the work addresses this algorithmic part and most often not in sufficient detail for an in-depth analysis. We hence extended the literature in this section, to draw from other related work to provide a better overview. Fourth, we asked how the impact of explainability is measured in the identified literature. We found that most of the literature looks at three aspects: (1) trust, (2) robustness, and (3) efficiency, of which trust and robustness have received the most attention. We looked at how these aspects are measured and formulated open questions for future work.

Looking at the big picture, we identified three strands of research. The first one has partial overlap with XAI and tries to implement decision support systems on embodied agents. The second treats humans as observers and investigates what kind of explainability mechanisms can be used to make humans understand the embodied agent's inner workings.

The third and final strand investigates how explainability mechanisms can be smoothly integrated into an interaction context (compared to treating humans as pure observers).

Finally, we provided a list of open questions and gaps in the literature that we identified during our analysis in the hope that further investigation will address this fascinating new domain of research.

ACKNOWLEDGMENT

This work received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 765955 (ANIMATAS Innovative Training Network).

REFERENCES

[1] Kumar Akash, Katelyn Polson, Tahira Reid, and Neera Jain. 2018. Improving Human-Machine Collaboration Through Transparency-based Feedback – Part I : Human Trust and Workload Model. *Elsevier* 51, 34 (2018), 315–321. https://doi.org/10.1016/j.ifacol.2019.01.028

[2] Kumar Akash, Tahira Reid, and Neera Jain. 2018. Improving Human-Machine Collaboration Through Transparency-based Feedback – Part II: Control Design and Synthesis. *Elsevier* 51, 34 (2018), 322–328. https://doi.org/10.1016/j.ifacol.2019.01.026

[3] Victoria Alonso and Paloma De La Puente. 2018. System transparency in shared autonomy: A mini review. , 83 pages. https://doi.org/10.3389/fnbot.2018.00083

[4] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable Agents and Robots: Results from a Systematic Literature Review. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '19)*. ACM, Montreal, 1078–1088. www.ifaamas.org

[5] Brenna D. Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. 2009. A survey of robot learning from demonstration. *Robotics and Autonomous Systems* 57, 5 (2009), 469–483. https://doi.org/10.1016/j.robot.2008.10.024

[6] M. Arnold, D. Piorkowski, D. Reimer, J. Richards, J. Tsay, K. R. Varshney, R. K.E. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilovic, R. Nair, K. Natesan Ramamurthy, and A. Olteanu. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63, 4-5 (2019), 6:1 – 6:13. https://doi.org/10.1147/JRD.2019.2942288 arXiv:1808.07261

[7] Kim Baraka, Stephanie Rosenthal, and Manuela M. Veloso. 2016. Enhancing human understanding of a mobile robot's state and actions using expressive Lights. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, New York, 652 – 657. https://doi.org/10.1109/ROMAN.2016.7745187

[8] Kim Baraka and Manuela M. Veloso. 2018. Mobile Service Robot State Revealing Through Expressive Lights: Formalism, Design, and Evaluation. *International Journal of Social Robotics* 10, 1 (jan 2018), 65–92. https://doi.org/10.1007/s12369-017-0431-x

[9] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012 arXiv:1910.10045

[10] Michael W. Boyce, Jessie Y.C. C. Chen, Anthony R. Selkowitz, and Shan G. Lakhmani. 2015. Effects of Agent Transparency on Operator Trust. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts (HRI'15 Extended Abstracts)*, Vol. 02-05-Marc. IEEE, New York, NY, USA, 179–180. https://doi.org/10.1145/2701973.2702059

[11] Joost Broekens and Mohamed Chetouani. 2019. Towards Transparent Robot Learning through TDRL-based Emotional Expressions. *IEEE Transactions on Affective Computing* -1, -1 (jan 2019), 1–1. https://doi.org/10.1109/taffc.2019.2893348

[12] Barry Brown and Eric Laurier. 2017. The trouble with autopilots: Assisted and autonomous driving on the social road. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, Vol. 2017-May. ACM, Denver, 416–429. https://doi.org/10.1145/3025453.3025462

[13] Tathagata Chakraborti, Anagha Kulkarni, Sarath Sreedharan, David E. Smith, and Subbarao Kambhampati. 2018. Explicability? Legibility? Predictability? Transparency? Privacy? Security? The Emerging Landscape of Interpretable Agent Behavior. arXiv:1811.09722 http://arxiv.org/abs/1811.09722

[14] Tathagata Chakraborti, Sarath Sreedharan, and Subbarao Kambhampati. 2019. Balancing explicability and explanations for human-aware planning. In *28th International Joint Conference on Artificial Intelligence (IJCAI)*, Vol. 2019-Augus. IJCAI, Macao, 1335–1343. https://doi.org/10.24963/ijcai.2019/185 arXiv:1708.00543

[15] Tathagata Chakraborti, Sarath Sreedharan, Anagha Kulkarni, and Subbarao Kambhampati. 2018. Projection-Aware Task Planning and Execution for Human-in-the-Loop Operation of Robots in a Mixed-Reality Workspace. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Madrid, 4476–4482. https://doi.org/10.1109/IROS.2018.8593830

[16] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. 2017. Plan Explanations as Model Reconciliation: Moving Beyond Explanation as Soliloquy. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*. Springer Nature, Melbourne, 156–163. arXiv:cs.AI/1701.08317

[17] Crystal Chao, Maya Cakmak, and Andrea L. Thomaz. 2010. Transparent active learning for robots. In *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction (HRI)*. IEEE, Osaka, 317–324. https://doi.org/10.1109/HRI.2010.5453178

[18] Jessie Y. C. Chen, Michael J. Barnes, Anthony R. Selkowitz, and Kimberly Stowers. 2017. Effects of Agent Transparency on human-autonomy teaming effectiveness. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, Budapest, 1838–1843. https://doi.org/10.1109/SMC.2016.7844505

[19] Jessie Y. C. Chen, Shan G. Lakhmani, Kimberly Stowers, Anthony R. Selkowitz, Julia L. Wright, and Michael J. Barnes. 2018. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science* 19, 3 (2018), 259–282. https://doi.org/10.1080/1463922X.2017.1315750

[20] Jessie Y. C. Chen, Katelyn Procci, Michael Boyce, Julia Wright, Andre Garcia, and Michael J. Barnes. 2014. *Situation Awareness–Based Agent Transparency*. Technical Report April. Army Research Laboratory, Washington, DC, USA. 1–29 pages. https://www.researchgate.net/publication/264963346{_}Situation{_}Awareness-Based{_}Agent{_}Transparency

[21] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. arXiv:1702.086ACM08 http://arxiv.org/abs/1702.08608

[22] Anca D Dragan, Kenton C.T. Lee, and Siddhartha S Srinivasa. 2013. Legibility and Predictability of Robot Motion. In *7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, Tokyo, 301–308. https://doi.org/10.1109/HRI.2013.6483603

[23] Mica R. Endsley. 2017. Toward a theory of situation awareness in dynamic systems. *Human Error in Aviation* 37 (2017), 217–249. https://doi.org/10.4324/9781315092898-13

[24] Heike Felzmann, Eduard Fosch-Villaronga, Christoph Lutz, and Aurelia Tamo-Larrieux. 2019. Robots and Transparency: The Multiple Dimensions of Transparency in the Context of Robot Technologies. *IEEE Robotics and Automation Magazine* 26, 2 (mar 2019), 71–78. https://doi.org/10.1109/MRA.2019.2904644

[25] Agneta H. Fischer and Antony S. R. Manstead. 2008. Social Functions of Emotion and Emotion Regulation. In *Handbook of emotions*. Guilford Press, unknown, Chapter 24, 456–468.

[26] Kerstin Fischer, Hanna Mareike Weigelin, and Leon Bodenhagen. 2018. Increasing trust in human-robot medical interactions: Effects of transparency and adaptability. *Paladyn* 9, 1 (2018), 95–109. https://doi.org/10.1515/pjbr-2018-0007

[27] Michael W. Floyd and David W. Aha. 2016. *Incorporating transparency during Trust-Guided behavior adaptation*. Vol. 9969 LNAI. Springer, Atlanta. 124–138 pages. https://doi.org/10.1007/978-3-319-47096-2_9

[28] Michael W. Floyd and David W. Aha. 2017. Using explanations to provide transparency during trust-guided behavior adaptation 1. *AI Communications* 30, 3-4 (2017), 281–294. https://doi.org/10.3233/AIC-170733

[29] Michael Georgeff, Barney Pell, Martha Pollack, Milind Tambe, and Michael Wooldridge. 1999. The belief-desire-intention model of agency. In *Intelligent Agents V: Agents Theories, Architectures, and Languages*, Vol. 1555. Springer, Paris, 1–10. https://doi.org/10.1007/3-540-49057-4_1

[30] Ze Gong and Yu Zhang. 2018. Behavior Explanation as Intention Signaling in Human-Robot Teaming. In *27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, Nanjing, 1005–1011. https://doi.org/10.1109/ROMAN.2018.8525675

[31] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L. Isbell, and Andrea Thomaz. 2013. Policy shaping: Integrating human feedback with Reinforcement Learning. In *Advances in Neural Information Processing Systems*, C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger (Eds.). NIPS, Lake Tahoe, 2625–2633. http://papers.nips.cc/paper/5187-policy-shaping-integrating-human-feedback-with-reinforcement-learning.pdf

[32] Elena Corina Grigore, Alessandro Roncone, Olivier Mangin, and Brian Scassellati. 2018. Preference-Based Assistance Prediction for Human-Robot Collaboration Tasks. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Madrid, 4441–4448. https://doi.org/10.1109/IROS.2018.8593716

[33] Jonathan Grizou, Manuel Lopes, and Pierre Yves Oudeyer. 2013. Robot learning simultaneously a task and how to interpret human instructions. In *2013 IEEE 3rd Joint International Conference on Development and Learning and Epigenetic Robotics, ICDL 2013 - Electronic Conference Proceedings*. IEEE, Osaka, 1–8. https://doi.org/10.1109/DevLrn.2013.6652523

[34] Bradley Hayes and Julie A. Shah. 2017. Improving Robot Controller Transparency Through Autonomous Policy Explanation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17*. IEEE, New York, New York, USA, 303–312. https://doi.org/10.1145/2909824.3020233

[35] High-Level Expert Group on Artificial Intelligence, AI HLEG, and High-Level Expert Group on Artificial Intelligence. 2019. Ethics Guidelines for Trustworthy AI. , 41 pages. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[36] Sandy H. Huang, David Held, Pieter Abbeel, and Anca D. Dragan. 2019. Enabling Robots to Communicate Their Objectives. *Autonomous Robots* 43, 2 (jul 2019), 309–326. https://doi.org/10.1007/s10514-018-9771-0 arXiv:1702.03465

[37] Xiangyang Huang, Cuihuan Du, Yan Peng, Xuren Wang, and Jie Liu. 2013. Goal-oriented action planning in partially observable stochastic domains. In *Proceedings - 2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems, IEEE CCIS 2012*, Vol. 3. IEEE, IEEE, Hangzhou, 1381–1385. https://doi.org/10.1109/CCIS.2012.6664612

[38] Giulio Jacucci, Anna Spagnolli, Jonathan Freeman, and Luciano Gamberini. 2014. Symbiotic interaction: A critical definition and comparison to other human-computer paradigms. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8820 (2014), 3–20. https://doi.org/10.1007/978-3-319-13500-7_1

[39] Mahdi Khoramshahi and Aude Billard. 2019. A dynamical system approach to task-adaptation in physical human–robot interaction. *Autonomous Robots* 43, 4 (apr 2019), 927–946. https://doi.org/10.1007/s10514-018-9764-z

[40] W. Bradley Knox and Peter Stone. 2009. Interactively Shaping Agents via Human Reinforcement: The TAMER Framework. In *Proceedings of the Fifth International Conference on Knowledge Capture (K-CAP '09)*. ACM, New York, NY, USA, 9–16. https://doi.org/10.1145/1597735.1597738

[41] W. Bradley Knox, Peter Stone, and Cynthia Breazeal. 2013. *Training a Robot via Human Feedback: A Case Study.* Lecture Notes in Computer Science, Vol. 8239. Springer, Bristol, Book section 46, 460–470. https://doi.org/10.1007/978-3-319-02675-6_46

[42] Alice Koller and John R. Searle. 1970. Speech Acts: An Essay in the Philosophy of Language. *Language* 46, 1 (1970), 217. https://doi.org/10.2307/412428

[43] Anagha Kulkarni, Satya Gautam Vadlamudi, Yantian Zha, Yu Zhang, Tathagata Chakraborti, and Subbarao Kambhampati. 2019. Explicable planning as minimizing distance from expected behavior. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS* 4 (2019), 2075–2077. https://dl.acm.org/doi/10.5555/3306127.3332015

[44] Minae Kwon, Sandy H. Huang, and Anca D. Dragan. 2018. Expressing Robot Incapability. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction.* IEEE, Chicago, 87–95. https://doi.org/10.1145/3171221.3171276

[45] Maurice Lamb, Riley Mayr, Tamara Lorenz, Ali A. Minai, and Michael J. Richardson. 2018. The Paths We Pick Together: A Behavioral Dynamics Algorithm for an HRI Pick-and-Place Task. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction.* IEEE, Chicago, 165–166. https://doi.org/10.1145/3173386.3177022

[46] Jin Joo Lee, W. Bradley Knox, Jolie B. Wormwood, Cynthia Breazeal, and David DeSteno. 2013. Computationally modeling interpersonal trust. *Frontiers in Psychology* 4, DEC (2013), 893. https://doi.org/10.3389/fpsyg.2013.00893

[47] Min Kyung Lee, Sara Kielser, Jodi Forlizzi, Siddhartha Srinivasa, and Paul Rybski. 2010. Gracefully mitigating breakdowns in robotic services. In *5th ACM/IEEE International Conference on Human-Robot Interaction, HRI 2010.* IEEE, Osaka, 203–210. https://doi.org/10.1145/1734454.1734544

[48] Michael S. Lee. 2019. Self-Assessing and Communicating Manipulation Proficiency Through Active Uncertainty Characterization. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction*, Vol. 2019-March. IEEE, Daegu, 724–726. https://doi.org/10.1109/HRI.2019.8673083

[49] Phil Legg, Jim Smith, and Alexander Downing. 2019. Visual analytics for collaborative human-machine confidence in human-centric active learning tasks. *Human-centric Computing and Information Sciences* 9, 5 (dec 2019), 1 – 25. https://doi.org/10.1186/s13673-019-0167-8

[50] Jamy Li. 2015. The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human Computer Studies* 77 (2015), 23–37. https://doi.org/10.1016/j.ijhcs.2015.01.001

[51] Zachary C. Lipton. 2018. The Mythos of Model Interpretability. *Queue* 16, 3 (jun 2018), 30:31—-30:57. https://doi.org/10.1145/3236386.3241340

[52] Ingo Lütkebohle, Julia Peltason, Lars Schillingmann, Britta Wrede, Sven Wachsmuth, Christof Elbrechter, and Robert Haschke. 2009. The curious robot - Structuring interactive robot learning. In *2009 IEEE International Conference on Robotics and Automation (ICRA2009).* IEEE, Kobe, 4156–4162. https://doi.org/10.1109/ROBOT.2009.5152521

[53] Joseph B. Lyons. 2013. Being Transparent about Transparency : A Model for Human-Robot Interaction. In *AAAI Spring Symposium Series.* AAAI, Stanford, 48–53. https://doi.org/10.1021/ja00880a025

[54] BF Malle. 2005. How the mind explains behavior: folk explanations, meaning, and social interaction. *Choice Reviews Online* 42, 10 (2005), 42–6170–42–6170. https://doi.org/10.5860/choice.42-6170

[55] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. , 38 pages. https://doi.org/10.1016/j.artint.2018.07.007 arXiv:1706.07269

[56] Anis Najar, Olivier Sigaud, and Mohamed Chetouani. 2016. Training a robot with evaluative feedback and unlabeled guidance signals. In *25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).* IEEE, Columbia, 261–266. https://doi.org/10.1109/ROMAN.2016.7745140

[57] Anis Najar, Olivier Sigaud, and Mohamed Chetouani. 2020. Interactively shaping robot behaviour with unlabeled human instructions. *Autonomous Agents and Multi-Agent Systems* 34, 2 (2020), 1 – 35. https://doi.org/10.1007/s10458-020-09459-6 arXiv:1902.01670

[58] Scott Ososky, Tracy Sanders, Florian Jentsch, Peter Hancock, and Jessie Y. C. Chen. 2014. Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems. In *Unmanned Systems Technology XVI*, Robert E. Karlsen, Douglas W. Gage, Charles M. Shoemaker, and Grant R. Gerhart (Eds.), Vol. 9084. Spie, unknown, 90840E. https://doi.org/10.1117/12.2050622

[59] Victor Paléologue, Jocelyn Martin, Amit K. Pandey, Alexandre Coninx, and Mohamed Chetouani. 2017. Semantic-based interaction for teaching robot behavior compositions. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).* IEEE, Lisbon, 50–55. https://doi.org/10.1109/ROMAN.2017.8172279

[60] Leah Perlmutter, Eric M. Kernfeld, and Maya Cakmak. 2016. Situated Language Understanding with Human-like and Visualization-Based Transparency. In *Robotics: Science and Systems*, Vol. 12. IEEE, Michigan, 40–50. https://doi.org/10.15607/rss.2016.xii.040

[61] Adam Poulsen, Oliver K. Burmeister, and David Tien. 2018. Care Robot Transparency Isn't Enough for Trust. In *2018 IEEE Region 10 Symposium, Tensymp 2018.* IEEE, Sydney, 293–297. https://doi.org/10.1109/TENCONSpring.2018.8692047

[62] Alessandro Roncone, Olivier Mangin, and Brian Scassellati. 2017. Transparent role assignment and task allocation in human robot collaboration. In *IEEE International Conference on Robotics and Automation (ICRA).* IEEE, Singapore, 1014–1021. https://doi.org/10.1109/ICRA.2017.7989122

[63] Avi Rosenfeld and Ariella Richardson. 2019. Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems* 33, 6 (nov 2019), 673–705. https://doi.org/10.1007/s10458-019-09408-y arXiv:1904.08123

[64] Kristin E. Schaefer, Ralph W. Brewer, Joe Putney, Edward Mottern, Jeffrey Barghout, and Edward R. Straub. 2016. Relinquishing manual control collaboration requires the capability to understand robot intention. In *International Conference on Collaboration Technologies and Systems (CTS)*. IEEE, Orlando, 359–366. https://doi.org/10.1109/CTS.2016.69

[65] Alessandra Sciutti, Laura Patanè, Francesco Nori, and Giulio Sandini. 2014. Understanding object weight from human and humanoid lifting actions. *IEEE Transactions on Autonomous Mental Development* 6, 2 (2014), 80–92. https://doi.org/10.1109/TAMD.2014.2312399

[66] A Carlisle Scott, William J Clancey, Randall Davis, and Edward H Shortliffe. 1977. Explanation Capabilities of Production-Based Consultation Systems. *American Journal of Computational Linguistics* 14, J77-1 (1977), 1–50. https://www.aclweb.org/anthology/J77-1006

[67] Sara Sheikholeslami, Justin W. Hart, Wesley P. Chan, Camilo P. Quintero, and Elizabeth A. Croft. 2018. Prediction and Production of Human Reaching Trajectories for Human-Robot Interaction. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, Chicago, 321–322. https://doi.org/10.1145/3173386.3176924

[68] Anna Spagnolli, Lily E. Frank, Pim Haselager, and David Kirsh. 2018. *Transparency as an Ethical Safeguard*. Vol. 10727. IEEE, Eindhoven. 1–6 pages. https://doi.org/10.1007/978-3-319-91593-7_1

[69] Halit B. Suay and Sonia Chernova. 2011. Effect of human guidance and state space size on Interactive Reinforcement Learning. In *20th IEEE International Symposium in Robot and Human Interactive Communication (Ro-Man 2011)*. IEEE, Atlanta, 1–6. https://doi.org/10.1109/ROMAN.2011.6005223

[70] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[71] Andreas Theodorou, Robert H Wortham, and Joanna J Bryson. 2016. Why is my Robot Behaving Like That? Designing Transparency for Real Time Inspection of Autonomous Robots. In *AISB Workshop on Principles of Robotics*. unknown, Sheffield, 4. http://opus.bath.ac.uk/49713/

[72] Andrea L. Thomaz and Cynthia Breazeal. 2006. Transparency and Socially Guided Machine Learning. In *5th International Conference on Development and Learning 2006 (ICDL06)*. IEEE, Bloomington, 1–146. https://dspace.mit.edu/handle/1721.1/36160https://www.cc.gatech.edu/fac/athomaz/papers/ThomazBreazeal-ICDL06.pdf

[73] Andrea L. Thomaz and Cynthia Breazeal. 2008. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence* 172, 6 (2008), 716–737. https://doi.org/10.1016/j.artint.2007.09.009

[74] William van Melle. 1978. MYCIN: a knowledge-based consultation program for infectious disease diagnosis. *International Journal of Man-Machine Studies* 10, 3 (1978), 313–322. https://doi.org/10.1016/S0020-7373(78)80049-2

[75] Ning Wang, David V. Pynadath, and Susan G. Hill. 2016. The Impact of POMDP-Generated Explanations on Trust and Performance in Human-Robot Teams. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS (AAMAS '16)*. ACM, Richland, SC, 997–1005. https://dl.acm.org/citation.cfm?id=2937071

[76] Ning Wang, David V. Pynadath, and Susan G. Hill. 2016. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Vol. 2016-April. IEEE, Christchurch, 109–116. https://doi.org/10.1109/HRI.2016.7451741

[77] Ning Wang, David V. Pynadath, Ericka Rovira, Michael J. Barnes, and Susan G. Hill. 2018. *Is it my looks? Or something i said? The impact of explanations, embodiment, and expectations on trust and performance in human-robot teams*. Vol. 10809. Springer, Waterloo. 56–69 pages. https://doi.org/10.1007/978-3-319-78978-1_5

[78] David Warren. 1977. *Implementing Prolog-Compiling Predicate Logic Programs*. Technical Report. Technical Report 39 and 40, Dept. of Artificial Intelligence, Univ. of Edinburgh., Edinburgh.

[79] Michael R. Wick and William B. Thompson. 1992. Reconstructive expert system explanation. *Artificial Intelligence* 54, 1-2 (1992), 33–70. https://doi.org/10.1016/0004-3702(92)90087-E

[80] A. William Evans, Matthew Marge, Ethan Stump, Garrett Warnell, Joseph Conroy, Douglas Summers-Stay, and David Baran. 2017. The future of human robot teams in the army: Factors affecting a model of human-system dialogue towards greater team collaboration. In *Advances in Human Factors in Robots and Unmanned Systems*, Vol. 499. Springer, Walt Disney World, 197–210. https://doi.org/10.1007/978-3-319-41959-6_17

[81] Robert H Wortham, Andreas Theodorou, and Joanna J Bryson. 2016. What Does the Robot Think? Transparency as a Fundamental Design Requirement for Intelligent Systems. In *IJCAI-2016 Ethics for Artificial Intelligence Workshop*. none, New York, New York, USA, –1. http://www.robwortham.com/instinct-planner/http://opus.bath.ac.uk/50294/1/WorthamTheodorouBryson{_}EFAI16.pdf

[82] Huidi Zhang and Shirong Liu. 2009. Design of autonomous navigation system based on affective cognitive learning and decision-making. In *2009 IEEE International Conference on Robotics and Biomimetics, ROBIO 2009*. IEEE, IEEE, Guilin, 2491–2496. https://doi.org/10.1109/ROBIO.2009.5420477

[83] Allan Zhou, Dylan Hadfield-Menell, Anusha Nagabandi, and Anca D. Dragan. 2017. Expressive Robot Motion Timing. In *ACM/IEEE International Conference on Human-Robot Interaction*, Vol. Part F1271. IEEE, New York, New York, USA, 22–31. https://doi.org/10.1145/2909824.3020221 arXiv:1802.01536

A IDENTIFIED CATEGORIES BY PAPER

Table 5. Identified Categories by Paper

CitationKey	Definition					Social Cues					Measurement				Learning Paradigm
	transparency	explainability	expressivity	other	none	Movement	Text	Speech	Imagery	other/None	Trust	Robustness	Efficiency	Other	ML
Akash et al. [1]	X	.	.	.	.	.	X	.	.	.	X	.	.	.	.
Akash et al. [2]	X	.	.	.	.	.	X	.	.	.	X	.	X	.	.
Baraka and Veloso [8]	X	.	X	.	.	.	.	.	.	X	X	X	.	.	.
Boyce et al. [10]	X	.	.	.	.	.	.	.	X	.	X	.	.	.	.
Brown and Laurier [12]	.	.	.	.	X	X	.	.	X	.	.	.	.	X	.
Chakraborti et al. [16]	.	X	.	.	.	.	.	.	.	X	X	.	X	.	.
Chao et al. [17]	X	.	.	.	.	X	.	.	.	.	X	X	X	.	X
Chen et al. [18]	X	.	.	.	.	.	X	.	.	.	X	X	X	.	.
Fischer et al. [26]	X	.	.	.	.	.	.	X	.	.	X	.	.	.	.
Floyd and Aha [27]	X	.	.	.	.	.	X	.	.	.	X	.	.	.	.
Floyd and Aha [28]	X	X	.	.	.	.	X	.	.	.	X	.	.	.	.
Gong and Zhang [30]	.	X	.	.	.	.	X	.	.	.	X	.	.	.	X
Hayes and Shah [34]	X	X	.	.	.	.	X	.	.	.	.	.	.	X	X
Huang et al. [36]	.	.	.	X	.	X	.	.	.	.	.	.	.	.	X
Khoramshahi and Billard [39]	.	.	.	.	.	.	.	.	.	X	.	.	.	.	X
Kwon et al. [44]	.	.	X	.	.	X	.	.	.	.	X	X	.	.	.
Lamb et al. [45]	.	.	.	.	X	X	.	.	.	.	.	X	.	.	.
Lee [48]	.	.	.	X	.	.	.	.	.	X	X	.	.	.	.
Legg et al. [49]	.	.	.	.	X	.	.	.	X	.	.	X	.	.	X
Lütkebohle et al. [52]	.	.	.	.	X	X	.	X	.	.	.	.	.	X	X
Perlmutter et al. [60]	X	.	.	.	.	X	.	X	X	X	.	X	X	.	X
Poulsen et al. [61]	X	.	.	.	.	.	X	.	.	.	X	.	.	.	.
Roncone et al. [62]	X	.	.	.	.	.	.	X	X	.	.	.	X	.	.
Schaefer et al. [64]	.	.	.	X	.	.	.	.	X	.	X	.	.	.	.

Table 5 continued from previous page

CitationKey	Definition					Social Cues					Measurement				Learning Paradigm
	transparency	explainability	expressivity	other	none	Movement	Text	Speech	Imagery	other/None	Trust	Robustness	Efficiency	Other	ML
Sciutti et al. [65]	.	.	.	X	.	X	.	.	.	.	.	X	.	.	.
Sheikholeslami et al. [67]	.	.	.	X	.	X	.	.	.	.	.	X	X	.	.
Chakraborti et al. [14]	.	X	.	.	.	.	X	.	.	.	.	X	.	.	.
Arnold et al. [6]	.	X	.	.	.	.	.	X	.	.	X	.	.	.	X
Wang et al. [75]	.	X	.	.	.	.	X	.	.	.	X	.	X	.	.
Wang et al. [76]	.	X	.	.	.	.	X	.	.	.	.	X	X	.	.
Wang et al. [77]	.	X	.	.	.	.	X	.	.	.	X	X	.	.	X
Zhou et al. [83]	.	.	X	.	.	X	.	.	.	.	X	.	.	.	.
Grigore et al. [32]	.	.	.	X	.	.	.	.	.	X	.	X	.	.	.

B DEFINITIONS USED BY CORE PAPERS

Table 6. Definition of Explainability Ordered by Publication Year

Year(s)	Author(s)	Definition	Motivation
2008	Fischer and Manstead [25]	Robot explanations of its own actions designed to make the process and robot behaviors and capabilities accessible to the user	Trust
2009	Lütkebohle et al. [52]	Structure verbal and non-verbal dialog to guide human actions	Machine Teaching, Predictability, Human-Robot Collaboration
2010	Chao et al. [17]	Revealing to the teacher what is known and what is unclear	Machine Teaching, Predictability
2013	Lee et al. [46]	Develop expectancy-setting strategies and recovery strategies to forewarn people of a robot's limitations and reduce the negative consequence of breakdowns	Robot acceptance
2014	Sciutti et al. [65]	Convey cues about object features (e.g., weight) to the human partner using implicit communication	Human-Robot Collaboration
2015	Boyce et al. [10]	Display transparency information (SAT model [18]) in the interface of an autonomous robot	Trust
2016	Wang et al. [75]	Generate explanations of the robot's reasoning, communicate uncertainty and conflicting goals	Trust, Teamwork
2016	Perlmutter et al. [60]	Communicate robot's internal processes with human-like verbal and non-verbal behaviors	Communication, Visualization, Control

Year(s)	Author(s)	Definition	Motivation
2016	Schaefer et al. [64]	Convey the robot's reasoning processes or intent, understanding the control allocation processes, and human engagement or reengagement strategies	Human-Robot Collaboration, Trust
2017	Floyd and Aha [27, 28]	Layer that allows the agent to explain why it adapted its behaviours	Trust, Adaptation
2017	Hayes and Shah [34], [6]	Autonomosly synthesize policy descriptions and respond to both general and target queries by human collaborators	Human-Robot Collaboration, Control, Debug
2017	Roncone et al. [62]	Transfer information to the human partner about its own [robot] internal state and intents	Trust, Proficiency, Confidence, Uncertainty, Introspection
2017	Chakraborti et al. [16], [14]	Robot's attempt to move the human's model to be in conformance with its own.	Communication, Human-Robot Collaboration, Impedance Mismatch
2017	Zhou et al. [83]	Communicate the robot's internal state though timing	Perceived Naturalness, Human's Learning (Task Understanding)
2017	Chen et al. [18, 19],[1, 2] 2018	Descriptive quality of an interface pertaining to its abilities to afford an operator's comprehension about an intelligent agent's intent, performance, future plans, and reasoning process [20].	Trust, Human's Workload
2018	Kwon et al. [44]	Express robot's incapability and communicate both what the robot is trying to accomplish and why the robot is unable to accomplish it	Robot's Acceptance, Human-Robot Collaboration

Year(s)	Author(s)	Definition	Motivation
2018	Baraka et al. [7]	Externalize hidden information of an agent. Express robot behaviors that have a specific communicative purpose.	Human-Robot Collaboration, Control, Communication
2018	Gong and Zhang [30]	Explaining robot behaviours as intention, or explicitly signalling the robot's intentions	Teamwork, Human-Robot Collaboration
2018	Lamb et al. [45]	Implement behavioral dynamics models based on human decision-making dynamics	Human-Robot Collaboration
2019	Legg et al. [49]	Establish a two-way collaborative dialogue on data attributions between human and machine and express personal confidence in data attributions	Active Learning, Human-Robot Collaboration
2019	Poulsen et al. [61]	Deciphering the behaviour of intelligent others [81], allowing 'inspection of thoughts'.	Ethical Decision-Making, Trust
2019	Huang et al. [36]	Communicate information to correctly anticipate a robot's behaviours in novel situations and building an accurate mental model of the robot's objective function	Prediction, Human-Robot Coordination
2019	Khoramshahi and Billard [39]	Ensure the robot's behaviours complies with human intention, adapting generated motions (i.e., the desired velocity) to those intended by the human user	Human-Robot Collaboration