Explainable agents and robots: Results from a systematic literature review

Robotics Track

Sule Anjomshoae
Umeå University
Umeå, Sweden
sule.anjomshoae@umu.se

Amro Najjar
Umeå University
Umeå, Sweden
amro.najjar@umu.se

Davide Calvaresi
HES-SO
Sierre, Switzerland
davide.calvaresi@hevs.ch

Kary Främling
Umeå University
Umeå, Sweden
kary.framling@umu.se

ABSTRACT

Humans are increasingly relying on complex systems that heavily adopts Artificial Intelligence (AI) techniques. Such systems are employed in a growing number of domains, and making them explainable is an impelling priority. Recently, the domain of eXplainable Artificial Intelligence (XAI) emerged with the aims of fostering transparency and trustworthiness. Several reviews have been conducted. Nevertheless, most of them deal with data-driven XAI to overcome the opaqueness of black-box algorithms. Contributions addressing goal-driven XAI (e.g., explainable agency for robots and agents) are still missing. This paper aims at filling this gap, proposing a Systematic Literature Review. The main findings are *(i)* a considerable portion of the papers propose conceptual studies, or lack evaluations or tackle relatively simple scenarios; *(ii)* almost all of the studied papers deal with robots/agents explaining their behaviors to the human users, and very few works addressed inter-robot (inter-agent) explainability. Finally, *(iii)* while providing explanations to non-expert users has been outlined as a necessity, only a few works addressed the issues of personalization and context-awareness.

KEYWORDS

Explainable AI; goal-based XAI; autonomous agents; human-robot interaction

ACM Reference Format:
Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable agents and robots: Results from a systematic literature review. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 11 pages.

1 INTRODUCTION

As humans rely more and more on complex artificial intelligence systems, providing explanations for their decisions to support an effective human-system interaction becomes increasingly important. This direction is confirmed by the ratification of recent General Data Protection Regulation (GDPR) law which underlines the right to explanations [12]. Therefore, designing transparent and intelligible technologies is becoming an impelling necessity.

For these reasons, research on eXplainable Artificial Intelligence (XAI) [36] gained significant momentum in the last years. Nevertheless, recent studies on the explanation methods mainly focused on the data-driven algorithms aiming to *interpret* the results of "black-box" machine learning mechanisms such as deep neural networks (DNN) [108]. This research line, pushed by the intriguing results of DNNs (e.g., a DNN mistakenly labeling a tomato as a dog [97]), aims to interpret, or provide a meaning for an obscure machine learning model whose inner-workings are otherwise unknown or non-understandable by the human observer. Thus, the majority of the recent studies and surveys focus on providing an overview on interpretability and explainability of data-driven algorithms [5, 24, 35, 72, 86]. Despite the fact that agents and robots are becoming pervasive in the daily-living of users for several applications such as training, e-health, and ambient intelligence, to the best of our knowledge, literature reviews on *explainable agency* (i.e. explaining the behavior of goal-driven agents and robots) are still missing. However, when interacting with these systems, humans have a tendency to suppose that they have "mental states" allowing to understand the rationale for their actions [47]. Therefore, in case the behavior of a given agent or robot is not explained, the user makes up an explanation that does not necessarily reflect the AI's internal stance. This increases the risk of self-deception and may degrade the quality of the interaction. To a certain extent, dangerous situations may arise, putting at risk the user safety. According to the recent literature [5, 76], explanations help users to increase confidence and trust, whereas misunderstanding the intentions of the intelligent system creates discomfort and confusion.

Contribution

This paper presents a Systematic Literature Review (SLR) aiming at providing a comprehensive overview of existing works on explainable agency for robots and intelligent agents. This helps understand how these systems tackled the problem of presenting human understandable explanations.

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

The rest of the paper is organized as follows: Section 2 sets the scope of the SLR and presents the explanation phases. Section 3 describes the review protocol applied in the present study. Section 4 analyzes the outcomes of the applied methodology and presents its results. Section 5 discusses a number of future research directions. Finally, Section 6 concludes the paper.

2 BACKGROUND AND DEFINITIONS

Explainable AI refers to an artificial intelligence whose actions, recommendations and the underlying causes to its decisions are understandable by humans.

After a period of relatively low activity, the domain XAI started to gain more attention in the recent years and a large body of research dealing with explanations and intelligibility was produced. However, since this research is carried out by researchers from different backgrounds and disciplines (e.g., machine learning, robotics, multi-agent systems, cognitive sciences), there are no shared definitions.

By gaining insights on how data-driven XAI (*i.e.* explaining black-box algorithms) and goal-driven XAI (*i.e.* explainable agency) approach the notions of explainability, this section aims at setting up the boundaries of this SLR. Furthermore, this section presents the explanation phases.

2.1 Explainability in Data-Driven Domain

In machine learning, explanations are often related to the concept of *interpretability*. Systems are *interpretable* if their operations can be understood by a human through introspection or explanation [5]. For instance, Choo and Liu [20] defined the interpretability of a deep learning model as identifying features in input layer which are responsible for the prediction result at the output layer. As its name indicates, data-driven XAI is about understanding of a decision of a "black-box" machine learning mechanism given the data used as an input [35]. For this reason, this branch of XAI is interested in comprehending how the available data led to a decision, and whether, given the data and specific circumstances, the machine learning mechanism can be remodeled to output the same decision [50].

Note that other works (e.g., [24, 66, 71]) in data-driven XAI relied on supplementary concepts such as *justification*, *transparency*, and *comprehensibility* to define explainability and interpretability for machine learning mechanisms. Yet, highlighting the nuances between these concepts used in data-driven XAI is beyond the scope of this article.

2.2 Explainable Agency

Explainable Agency refers to the autonomous agents (e.g., robots) explaining their actions and the reasons leading to their decisions [60].

To understand the actions of robots and intelligent agents, humans tend to *mentalize* their behavior. Mentalizing or mindreading is based on the Theory of Mind (ToM) which postulates that humans estimate the actions of other humans by observing their behavior and attributing mental states (e.g., beliefs, desires, emotions, intentions, etc.) thereby attempting to understand their own perspective [32, 47, 80]. Using this ToM of others, humans are able to *(i)* overcome the complexity of the world since the ToM makes it possible to comprehend the behaviors of other people and avoid confusion, and *(ii)* predict the future behavior of others and deal with it [47, 69]. As it has been confirmed by recent research in the domain of human-robot interaction [61], humans tend to apply this ToM not only to other humans but also to non-human objects and robots. This tendency to anthropomorphize objects is known as the "intentional stance" [23]. This implies that, with the lack of explanation, the human user may construct an erroneous explanation of the robot or agent.

We define goal-driven XAI[1] as a research domain aiming at building explainable agents and robots capable of explaining their behavior to a lay user. These explanations would help the user to build a ToM of the intelligent agent and would lead to better human-agent collaboration and incite the user to understand the capabilities and the limits of the agents, thereby improving the levels of trust and safety, and avoiding failures, since the lack of appropriate mental models and knowledge about the agent may lead to failed interactions [4, 18].

As will be discussed in Section 3, this article reviews existing works about explainable agency defined above. Note that in the literature, and within the domain of social robots in particular, other related terms are also in use, including understandability [47], explicability[2] [16], transparency [105], predictability [26], readability [98], and legibility [64]. Defining these terms, comparing them and highlighting their similarities and differences are beyond the scope of this article[3]. Furthermore, surveying data-driven XAI (*i.e.*, works aiming to open black-box ML mechanisms) is also beyond the scope of this article. For this type of surveys, please refer to [1, 35, 86].

2.3 Explanation Phases

Aiming to integrate the user into the loop, and to address issues related to explanation communications, the authors in [74] distinguish three explanation phases:

(1) **Explanation Generation**: The aim is to generate an explanation justifying why an action/result was taken/achieved. The actual implementation of this phase is determined by the AI model of the agent (e.g. BDI agent [83]). Citing goals [8], desires [51] and emotions [52] are examples of the explanation generation process of the literature.
(2) **Explanation Communication**: Given the explanation generated in the previous phase, this phase deals with what exactly to be provided to the end-user and how to present it [74].
(3) **Explanation Reception**: This phase studies how well the human understands the explanation. To assess this, typically, research relies on user studies, subjective evaluation. Furthermore, in order to better understand explanation reception, meaningful metrics should be devised to assess the explanation and poll the users about it.

[1] In this article we use the terms Explainable Agency and Goal-Driven XAI interchangebly.
[2] In contrast to explainable agent where explicit explanation is required, [94] defines *explicabile* system as one that avoid the need to provide explanations by generating plans that match the human's expected plan.
[3] Please refer to [15] for further insights.

3 REVIEW METHODOLOGY

This study has been performed as a Systematic Literature Review (SLR). By doing so, it is possible to rigorously reproduce the retrieval, selection, and analysis processes of the relevant literature. This paper adheres to the procedure adopted and adapted by [10] (see Figure 1).

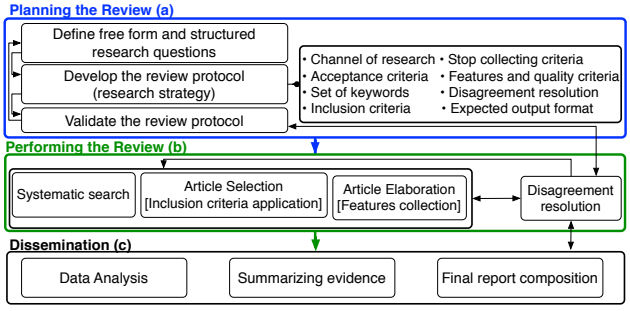

Figure 1: Review methodology adapted from [11] and [54].

Following the Goal-Question-Metric (GQM) [53], the generic free-form question *"What does imply having Explainable Agency (EA) in user- and system-centric domains?"* is broken-down in the following structured research questions (SRQs).

SRQ1: *Demographics* - How has EA been evolving over the years in terms of research domains? [e.g., when (year) and where (the geographical indication of the scientific institute)]

SRQ2: *Application scenarios* - What kinds of application scenarios have been addressed by the primary studies? [e.g., e-health, domestic robots, and training]

SRQ3: *Drives (needs)* - Which are the main drives demanding explainability?

SRQ4: *Social Science and psychological background* - How have the explanations been grounded onto the social-science background to provide understandable explanations?

SRQ5: *Design* - Which platforms and architectures have been used to design EAs? [e.g., BDI, MDP, POSH, etc.]

SRQ6: *Dynamics (Context-aware, user-aware)* - What explanatory granularity has been provided with respect to the user and its context?

SRQ7: *Presentation* - How have the explanations been presented for human-system interaction? [e.g., expressive lights, graphical user interface, natural language]

SRQ8: *Evaluation/ Framework* - How have the validity and utility of the explanations been evaluated by the authors of the primary studies?

SRQ9: *Future challenges* - What are the stated future research directions and challenges identified by the scientific community?

To perform a more accurate semi-automatic research, some keywords have been contextualized (keeping some keywords fixed in the performed queries). Based on the reviewers' rooted backgrounds on MAS, XAI, and robotics domains, the following keywords have been defined: **Explainable AI;** (Explainable AI + agent), (Explainable AI + robot), (Explainable AI + Transparency), **Agent;** (Intelligent agents + explanation), (Self-explaining agent + AI), (Explainable multi-agent systems), (Agent teamwork + explanation), (Understandable + agent), (Agent + explain + transparency), and **Robot;** (Human-robot interaction + readability), (Human-robot interaction + intention + legibility), (Human-robot interaction + interpretability), (Human-robot + intention recognition + prediction), (Understandable robot), (Explainable planning + robot), (Robot transparency).

The research of the articles has been conducted using the following sources: IEEExplore, Science Direct, ACM, and Google Scholar. Initially, 303 papers have been collected. On turn, they have been reduced to 62 papers[4] by performing a further coarse-grained, then a fine-grained examination. To do so, the abstracts and the text of the collected papers have been verified to comply with following the inclusion criteria:

A) Recent Paper (2008 or after): Since the aim is to identify the current trends and understand recent works addressing explainable agency, we chose to restrain this work to papers published in the last decade (2008-2018).

B) Relevance: The paper must be relevant for the XAI domain. For instance, papers addressing explanations in social science without any relevancy to AI are excluded.

C) Primary Study: Only papers providing a direct contribution on XAI (e.g., models, architectures, or implementations) are included, secondary studies (i.e. surveys) are excluded.

D) Accessibility: To be included, the content of the article should be accessible via one of the portals mentioned above.

E) Explainable Agency: The aim of this SLR is to study the explainability of goal-driven robots & agents (i.e., *cognitive explainability*, c.f. Section 2 and [74]). Data-driven XAI (i.e., machine learning interpretability) is beyond the scope of this article. Note that goal-driven agents/robots who rely on ML mechanisms (e.g. reinforcement learning) to update their knowledge about the user or the environment are included.

F) Singularity/Originality: Duplicate papers, or papers which have been published in an extended or complete version are not included. Only the complete version is included.

G) Explanation as a *Communicative Action*: Explanations should be provided by explicit communicative action. The latter is defined as an action performed by an agent/robot, with the intention of increasing another agent's knowledge of the first agent/robot [47, 55, 56]. Therefore, papers proposing *explicable* robots (c.f. Section 2.2) are excluded from this SLR since in these works, the robot does not communicate an explicit explanation to its users. Furthermore, note that the communicative action defined above can be a natural language message, images on a graphical user interface, or expressive lights (e.g., lights used to express the robots behavior [93]).

4 RESULT PRESENTATION

This section presents the results from the qualitative analysis of the studied papers.

[4]Table 1 lists the 62 papers retained for the SLR.

4.1 SRQ1: Demographics

The chronological and geographical distribution of works on explainable agency are shown in Figure 2 and Figure 3. Although there has been uneven proportion in the number of studies in early 2010s and before, there is an increasing growth over the last five years. This might be due to the effect of general emphasis on explainable AI and the "right to explanation" by the GDPR [101] and similar initiatives [12]. The trend may exponentially increase in upcoming years with the results of different research domains working together on XAI.

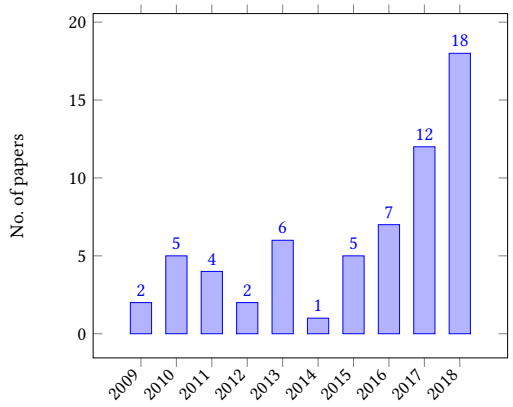

Figure 2: Number of papers per year.

To understand the geographic distributions of the research institutions working on the EA domain, the number of papers per country is plotted in Figure 3. The results show that most of the research institutions are based in the USA, followed by the Netherlands and the UK. It is also worth to note that the collaborative works among institutions have increased the number of papers in the aforementioned countries. The analysis also shows that the reviewed papers were written by authors from 26 different institutions. With the exception of the Netherlands, the number of articles published by European countries is relatively low. The number of papers published by the US researchers (*i.e.*, 33 publication) is higher than that of European countries combined. However, in the near future, the European research on this subject might increase with the promulgation of the GDPR.

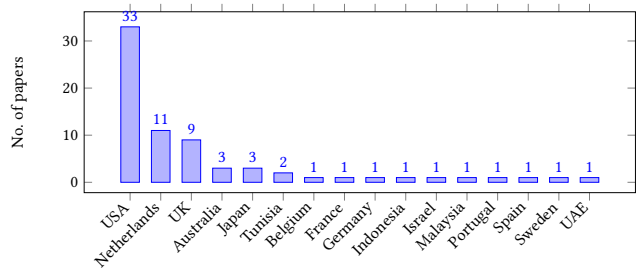

Figure 3: Number of papers per country.

Figure 4 presents the research domains working on EA. The Figure shows that the most active research area is the human-agent interaction, followed by human-robot collaboration, human-robot interaction, and human-agent collaboration. The difference between human-robot and human-agent collaboration studies is that the former one involves embodied robots while the latter concerns virtual intelligent systems. The results show that providing explanations were particularly emphasized in mixed autonomy settings where humans collaborate with autonomous agents or robots.

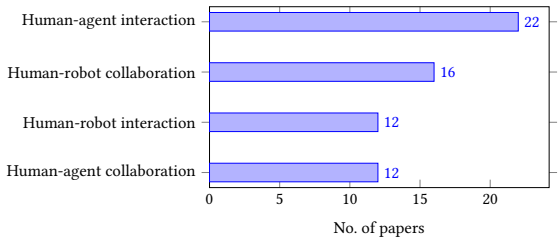

Figure 4: Research domains working on EA.

4.2 SRQ2: Application Scenarios

This section identifies the application scenarios in which explanations have been offered. As presented in Figure 5, the proposed application scenarios include; Robot collaborative task (29%), Robot navigation (20%), Game applications (17%), Search and rescue (10%), Training (9%), Recommender systems (5%), E-health (5%), and Ubiquitous computing (5%). Robot collaborative tasks such as working in a factory environment with humans (e.g., [44]) and teaming for military missions [104] are the particularly preferred cases of the surveyed articles. In game applications, explanations were provided for the non-player characters to reduce the frustration of the human players [73]. The educational explainable agents were evaluated in virtual a firefighting agent to train the crisis management team (e.g., [38]). In e-health, explanations were presented in a personal health assistant to help children to cope with type 1 diabetes (e.g., [52]). The other studies assessed explanations in ubiquitous computing to allow users to understand the system's reasoning (e.g., [100]), search and rescue scenarios for robot behavior (e.g., [67]), and movie and music recommender systems (e.g., [59]).

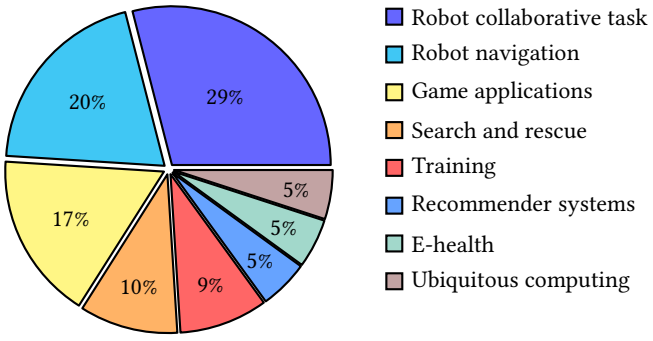

Figure 5: Application scenarios.

Table 1: The drives of explanations of the primary studies covered by the review.

Drives	Primary studies
Transparency	[89] [100] [81] [79] [103] [44] [106] [19] [95] [73] [41] [51] [107] [88] [14] [65] [43] [82] [34] [57] [29] [6] [7] [42] [104] [46] [18] [75]
Trust	[102] [58] [49] [100] [59] [29] [67] [6] [99] [52] [57] [27] [7] [42] [104] [31] [14] [65] [43] [62] [46] [29] [96]
Collaboration	[22] [78] [74] [63] [33] [62] [77] [31] [92] [102] [62] [14] [25] [44] [19] [37] [75]
Intent communication	[3] [98] [85] [13] [70] [25] [92] [93] [31] [78] [17] [52]
Control	[2] [34] [17] [95] [88] [82] [49] [100]
Education	[38] [39] [8] [40]
Debugging	[48] [91] [103]

4.3 SRQ3: Drive Demanding Explainability

Studies in the literature underlined the importance of considering the intended purpose when incorporating explanation facilities into intelligent systems [45]. This SRQ seeks to understand the reasons why the reviewed studies provided explanations. As a matter of fact, most of these works stated their motivation or intended purpose for the explanations one way or another. Table 1 lists the drives of the 62 papers included in the SLR. Note that some papers may have more than one drive.

Increasing user's trust in the system, transparency i.e. explaining the inner-workings of the systems to the user and informing about the intents of the agent (intent communication) are among the listed motivations for the explanations. The table reveals that trust, and transparency are the most prominent drives of the explanation. The studies show that transparency and trust are going hand in hand to increase the user's confidence in the system by understanding how its reasoning mechanism works (e.g., [14, 102]). In applications requiring human-robot interaction, intent communication is one of the main drives for explanations in order to make the robot's internal state (e.g. goals & intentions) understandable to humans [3]. For collaborative tasks, explanations were deemed essential to increase efficiency and team performance [63]. Explanations are also useful for control purposes. In particular, they were presented to determine the level of autonomy to grant to an agent (e.g., [2, 49]). Education and debugging were identified as the motivations of explanations, the former is referred to as allowing users to learn something from the system [52] and the latter is considered for notifying users about the defects in the system [48].

4.4 SRQ4: Social Science and Psychological Background

This research question explores the literature to find out whether the studied works were grounded on any social-science or psychological backgrounds. 39 of the studied papers did not rely on any theoretical background related to generating explanations. Figure 6 shows the relevant background adopted by the rest of the papers. As shown in the figure, the most cited social science theory was the folk psychology (30%) followed by theory of mind (27%), social psychology (7%), color psychology (7%), the hierarchical task analysis (HTA) (7%), and others include; proxemics, signal detection theory and rationalization. Folk psychology [21] refers to explaining human behaviors in terms of its underlying mental states such as beliefs, desires, and intentions.

While folk psychology can be considered as one type of Theories of Mind (ToM), papers listed under the ToM [87] category are those who relied on the general concept of ToM. Color psychology, used in expressive robots, is about investigating the various aspects of color, including color vision, color symbolism and association, and color effects on psychological and biological functioning [28]. HTA is a technique for cognitive task analysis to identify complex human tasks. This technique has been used in modeling agent's goals, beliefs and actions hierarchy [39]. The philosophy of language explores the nature of explanations and their relationship with linguistics [9]. Other workers relied on; proxemics [84], rationalization [27], and signal detection theory (Psychophysics) [68].

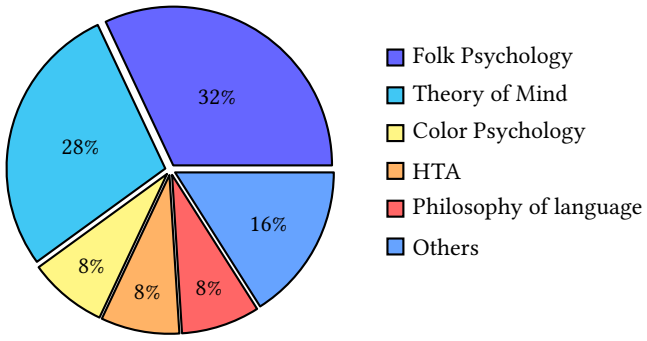

Figure 6: Social science and psychological background of the explanation methods.

4.5 SRQ5: Design of the Explanations

Figure 7 shows the platforms and architectures used to design the explainable agency. As can be seen from the figure, the majority of the works has not explicitly expressed their method for generating explanations. The result also shows that a number of papers relied on customized methods to address their own explanations problem (18). Following that, BDI (Belief, Desires, and Intentions) architecture (9) was widely implemented to generate explanations for goal-directed agents (e.g., [8, 74]). The rest of the platforms and architectures referenced to extract explanations are; Markov Decision Process (MDP) (3), Neural Networks (NN) (3), Partially Observable Markov Decision Process (POMDP) (3), Parallel-rooted-ordered Slip-stack Hierarchical Action Selection (POSH) (2), and Stanford Research Institute Problem Solver) (STRIPS) (2). While MDP, NN, and POMDP are employed for generating explanations of agent's goal and action, POSH and STRIPS are utilized to understand the robot behavior.

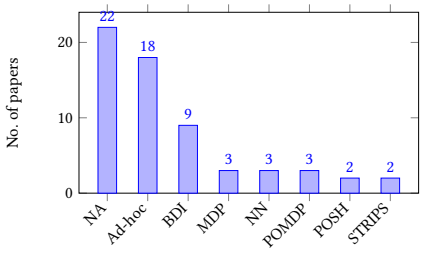

Figure 7: The used platforms and architectures.

4.6 SRQ6: Dynamics of the Explanation Method

This section explores what level of explanations has been provided with respect to the user and the context. As shown in Figure 8, while 41% of the studies have not addressed any of the given aspects, 43% of them provided context-aware explanations, 10% is user-aware, and in 6% of studies, explanations are tailored to both needs. Context-aware explanations generally refer to agents/robots that consider the context when selecting the best explanation to provide. Context-aware explanations have been proposed to implement effective control in ubiquitous systems (e.g., [65]), to facilitate context-aware explanations in human-robot teaming (e.g., [33]), and enhance robot navigation (e.g., [29, 43]). User-aware explanations are usually concerned with customizing the explanation based on the intended user, such as providing explanations depending on the user age [51], or personalized recommendations based on the user preferences [82]. The results show that relatively little research has been conducted for personalized explanations.

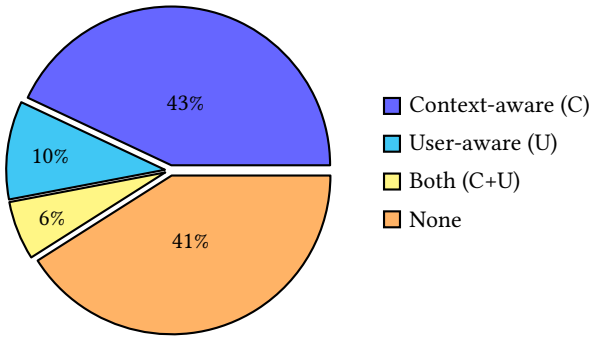

Figure 8: Dynamics of the explanation study.

4.7 SRQ7: Presentation of the Explanations

This section studies the types and categories of the explanations. The details of the analysis are discussed below.

The types of the explanations: As shown in Figure 9, six different presentation types are extracted from the 62 primary studies. The most frequent type is the text-based explanations (47%) and followed by visualization (21%), logs (11%), expressive motions (11%), expressive lights (7%), and speech (3%). Text-based explanations are presented in the form of natural language processing and traces (e.g., [41, 102]). Visuals such as graphs and images are integrated into the user interfaces to illustrate the explanatory information (e.g., [19, 65]). In human-robot interaction, expressive motion and expressive lights are stated as the most effective way of communicating the internal state of the robot [3]. While, some studies proposed speech as an alternative way to express explanation [62], generally the multimodal communication of explanations is under-represented.

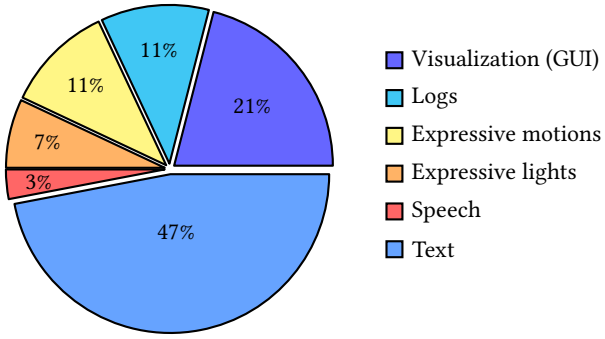

Figure 9: Types of explanation presentation.

The categories of the explanations: This analysis shows the quantification of papers based on the categories of explanations proposed by [90]. This analysis is important to assess the contexts of the explanations at a more abstract level. As demonstrated in Figure 10, the most prominent category is the *(i)* introspective informative explanations (26). This type of explanations is based on the reasoning process which leads to a decision to improve the interaction quality in human-agent/robot context. *(ii)* Teaching explanations (13) are the second commonly used explanation type which often aims to teach humans about concepts that agents/robots have learned. *(iii)* Introspective tracing explanations (10), similar to informative explanations, provide information about the underlying cause of a decision by giving more details about it. This type is often used to control the agent's behavior since they allow to discover the origin of a problem or to clarify misunderstandings between the system and the user. *(iv)* Execution explanations (9) report the list of operations the system undertakes. Finally, *(v)* post-hoc explanations (4) give explanations without necessarily tracing the actual reasoning process that led to the decision. Beside these explanation categories,a number of studies has provided *contrastive explanations* (e.g., [17, 74, 96]). These explanations refer to why a certain action or decision was chosen instead of another.

4.8 SRQ8: Evaluation of the Explanations

To answer this research question, we identify papers containing evaluations aiming to assess the validity and utility of the explanations. In the studied papers, most of the works either lack evaluations or conduct a user study for relatively simple scenarios. As presented in Figure 11, while 32% of the studies have not attempted any type of evaluations, 59% of them evaluated the usefulness and naturalness of the explanations through a user study. Some research assessed whether providing explanations has a positive effect on team performance and increase trust to system [63]. 9% of the

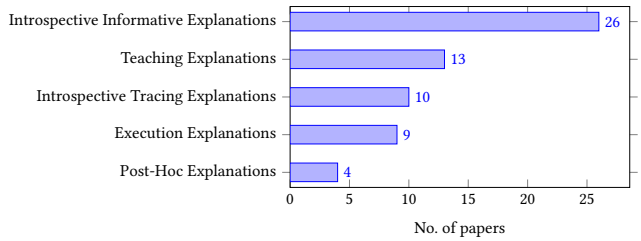

Figure 10: Categories of the explanations.

studies preferred empirical algorithmic evaluation using measures devised for their explanation problems. This might be due to limitations in time and subject availability to conduct a user study.

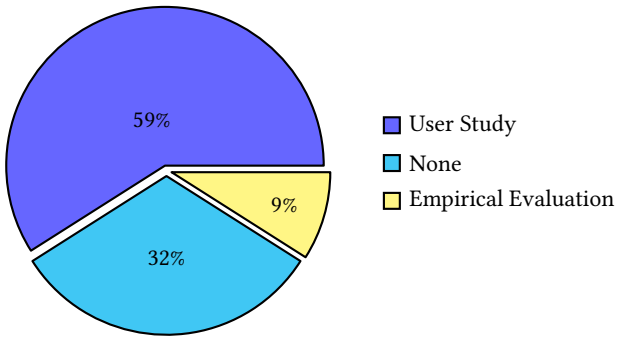

Figure 11: Evaluation of the explanations.

4.9 SRQ9: Future Challenges

To answer this research question, we clustered the future directions and challenges stated in the studied papers (see Figure 12). The majority of the papers pointed out the communication of the explanations (29%) as their future work. Other challenges and limitations mentioned in the literature were; conducting evaluations (20%), issues related to the core AI running the system (19%), context-awareness (14%), personalization (8%), emotions (4%), etc.

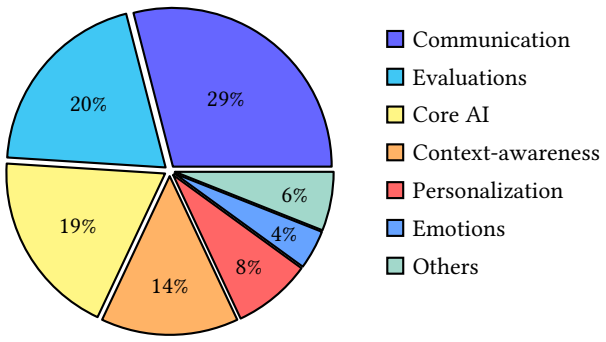

Figure 12: Future challenges stated by the primary studies.

5 DISCUSSION

The results presented in Section 4 unveiled the drives for explanations, as well as how they are generated, communicated, and evaluated. However, assessing the quality of the collected features enriches the overall picture.

5.1 Quality Criteria Assessment

The use of quality criteria assessment to report the quality of the primary studies is a well-established approach in SLR [11, 30]. The list of quality criteria defined in this work are: *(i)* how well the authors **motivated** their paper, *(ii)* details about the **context** and the design of the study, *(iii)* the quality of the statement of the major **results** and their analysis, *(iv)* to the extent to which the authors identified and discussed the **limitations** of their study. Figure 13 shows the quality of the 62 papers evaluated based on these criteria. Reviewers have graded the papers, according to their knowledge in the field, as three levels: "good", "arguable" and "poorly presented". Each paper was evaluated at least by 2 reviewers and their results averaged.

Overall, the motivation of the paper and the research gap are well defined in the studied papers. A number of studies lacked the clear content description and the presentation of their explanation methods. As can be seen from Figure 13, the result criteria have relatively low grade since a considerable portion of the studies has not reported any evaluation results. Generally, the limitations of the proposed methods are vaguely described and future research directions are not well defined in the reviewed papers. The results of this qualitative assessment confirm that research on explainable agents and robots is in its early stages of development. The next section offers a road-map for researches interested in this area.

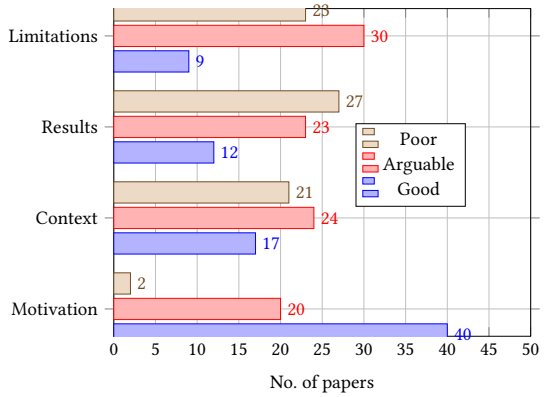

Figure 13: The qualitative assessment of the studied papers.

5.2 Phases of Explanation: A Detailed Representation

Figure 14(a) shows the three classic phases of EA (c.f. Section 2.3 & [74]). Nevertheless, there is a need for a more detailed representation to identify and organize the contributions and needed interventions (see Figure 14(b)). In Figure 14(b), *(i)* the *generation* of

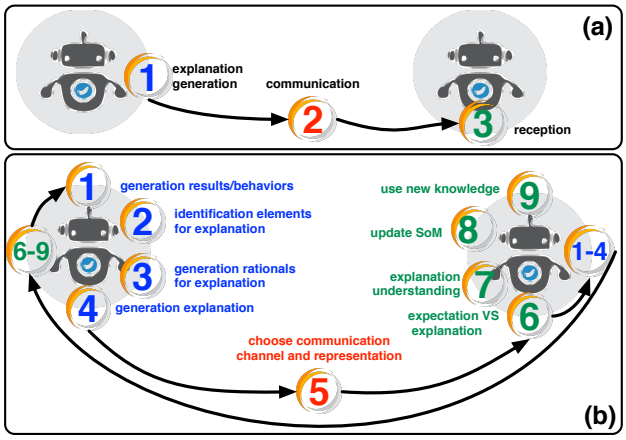

Figure 14: Phases of the explanation.

the explanation is composed of the tasks (1-4), *(ii)* the *communication* (5) concerns decisions and technologies about communication channels and ways of presenting the explanation, and the tasks (6-9) are about *(iii)* the *reception*. Furthermore, according to our investigations a final loop connecting the *reception* and the *generation* of an explanation must be added. This is due to the evidence that providing an explanation might require more than a single one-way interaction. Based on these findings, and according to the classification proposed in Figure 14(b), we present a road-map for explainability in the context of autonomous agents/robots, multi-agent systems and we suggest future strategic application domains.

5.3 Road-Map

This section proposes the envisioned research road-map based on the representation given in Figure 14(b).

5.3.1 Explanation Generation. This phase is tightly related with the core AI running the agent/robot. Based on the tasks identified in Figure 14, the key research directions for this phase are the following:

(1) Existing works present considerable advances in agent and robot architectures (e.g., cognitive architecture, and BDI architecture). Such architectures have an elaborate decision loop typically decomposed in several modules. However, most of them do not support explainability functions. To further push the research of EA, linking the agent/robot inner AI mechanism with the explanation generation module is a crucial step.

(2-4) While context-awareness and personalization have been outlined as key factors for EA [51], according to the results of this SLR, a few works in the literature address such issues. Thus, to generate dynamic explanation, there is a need for new mechanisms allowing the identification of relevant elements for an explanation (2), identifying its rationales (3), and integrating these elements into a sound explanation (4).

5.3.2 Explanation Communication. In this phase the *communicative act* [47, 55, 56] of explanation takes place, thereby sending the explanations to the user or to another agent. The identified steps follow.

(5) Explainable agents/robots are likely to be deployed in different types of environments. For this reason, the multi-modal explanation presentation (e.g., visual, audio, expressive) is a promising explanation communication approach. Although this approach is almost absent in the literature (as discussed Section 4.7), it is a promising research direction enabling an efficient EA communication. Yet, in such settings the agent/robots must be able to choose the communication channel and the representation. Note that this does not mean that all EA requires expensive and multi-modal explanation communication capacities. In some cases, it is possible to signal lots of information in cheap and small channels. For instance, using expressive lights for domestic robots [93].

5.3.3 Explanation Reception. A possible aim of the explanation is to make the receiver understand the State of Mind (SoM) of the sender. To ensure an accurate reception, the following points should be investigated:

(6-7) Metrics should be devised to assess how efficient the explanation is and how the user reacts to it. Such metrics can be relevancy, clarification, etc.

(8-9) The agent/robot should keep track of a model of the user knowledge. This model should be updated to reflect the evolution of the user expertise and how the user views the SoM of the agent/robot.

6 CONCLUSIONS

Driven by a growing need for, and interest in, transparent AI systems, this paper presented a SLR to clarify, map and analyze the relevant literature on explainable agents and robots in the last ten years. Alongside with the results presentation, the quality and the need for a more detailed representation of the explanation phases have been discussed. Moreover, a qualitative assessment of the studied papers has been reported. Finally, connected to the information elicited by this study, a road-map has been proposed to consolidate and guide new researchers who would like to tackle this field.

ACKNOWLEDGMENTS

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. Moreover, the authors would like to thank Ali Anjomshoae, Timotheus Kampik, and Yazan Mualla for their support during the camera-ready process.

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* (2018).

[2] Dan Amir and Ofra Amir. 2018. HIGHLIGHTS: Summarizing Agent Behavior to People. In *Proc. of the 17th International conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*.

[3] Kim Baraka, Ana Paiva, and Manuela Veloso. 2016. Expressive lights for revealing mobile service robot state. In *Robot 2015: Second Iberian Robotics Conference*. Springer, 107–119.

[4] Cindy L. Bethel. 2009. *Robots Without Faces: Non-verbal Social Human-robot Interaction.* Ph.D. Dissertation. Tampa, FL, USA. Advisor(s) Murphy, Robin R. and Hall, Lawrence O. AAI3420462.

[5] Or Biran and Courtenay Cotton. 2017. Explanation and justification in machine learning: A survey. In *IJCAI-17 Workshop on Explainable AI (XAI)*. 8.

[6] Rita Borgo, Michael Cashmore, and Daniele Magazzeni. 2018. Towards Providing Explanations for AI Planner Decisions. *arXiv preprint arXiv:1810.06338* (2018).

[7] Michael W Boyce, Jessie YC 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*. ACM, 179–180.

[8] Joost Broekens, Maaike Harbers, Koen Hindriks, Karel Van Den Bosch, Catholijn Jonker, and John-Jules Meyer. 2010. Do you get it? User-evaluated explainable BDI agents. In *German Conference on Multiagent System Technologies*. Springer, Springer, 28–39.

[9] Sylvain Bromberger. 1992. *On what we know we don't know: Explanation, theory, linguistics, and how questions shape them.* University of Chicago Press.

[10] Davide Calvaresi, Kevin Appoggetti, Luca Lustrissimini, Mauro Marinoni, Paolo Sernani, Aldo Franco Dragoni, and Michael Schumacher. 2018. Multi-Agent Systems' Negotiation Protocols for Cyber-Physical Systems: Results from a Systematic Literature Review. In *Proceedings of ICAART*.

[11] D. Calvaresi, D. Cesarini, P. Sernani, M. Marinoni, A.F. Dragoni, and A. Sturm. 2016. Exploring the ambient assisted living domain: a systematic review. *Journal of Ambient Intelligence and Humanized Computing* (2016), 1–19.

[12] Peter Carey. 2018. *Data protection: a practical guide to UK and EU law.* Oxford University Press, Inc.

[13] Ravi Teja Chadalavada, Henrik Andreasson, Robert Krug, and Achim J Lilienthal. 2015. That's on my mind! robot to human intention communication through on-board projection on shared floor space. In *Mobile Robots (ECMR), 2015 European Conference on*. IEEE, 1–6.

[14] Tathagata Chakraborti, Kshitij P Fadnis, Kartik Talamadupula, Mishal Dholakia, Biplav Srivastava, Jeffrey O Kephart, and Rachel KE Bellamy. 2018. Visualizations for an Explainable Planning Agent.. In *IJCAI*. 5820–5822.

[15] 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 preprint arXiv:1811.09722* (2018).

[16] Tathagata Chakraborti, Sarath Sreedharan, and Subbarao Kambhampati. 2018. Explicability versus Explanations in Human-Aware Planning. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2180–2182.

[17] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. 2017. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. *arXiv preprint arXiv:1701.08317* (2017).

[18] Arjun Chandrasekaran, Deshraj Yadav, Prithvijit Chattopadhyay, Viraj Prabhu, and Devi Parikh. 2017. It Takes Two to Tango: Towards Theory of AI's Mind. *arXiv preprint arXiv:1704.00717* (2017).

[19] Jessie YC Chen, Shan G Lakhmani, Kimberly Stowers, Anthony R Selkowitz, Julia L Wright, and Michael Barnes. 2018. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical issues in ergonomics science* 19, 3 (2018), 259–282.

[20] Jaegul Choo and Shixia Liu. 2018. Visual analytics for explainable deep learning. *IEEE computer graphics and applications* 38, 4 (2018), 84–92.

[21] Paul M Churchland. 1989. Folk psychology and the explanation of human behavior. *Philosophical Perspectives* 3 (1989), 225–241.

[22] Dustin Dannenhauer, Michael W Floyd, Matthew Molineaux, and David W Aha. 2018. Learning from Exploration: Towards an Explainable Goal Reasoning Agent. (2018).

[23] Daniel Clement Dennett. 1989. *The intentional stance.* MIT press.

[24] Derek Doran, Sarah Schulz, and Tarek R Besold. 2017. What does explainable AI really mean? A new conceptualization of perspectives. *arXiv preprint arXiv:1710.00794* 2071 (2017).

[25] Anca D Dragan, Shira Bauman, Jodi Forlizzi, and Siddhartha S Srinivasa. 2015. Effects of robot motion on human-robot collaboration. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 51–58.

[26] Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. 2013. Legibility and predictability of robot motion. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*. IEEE Press, 301–308.

[27] Upol Ehsan, Brent Harrison, Larry Chan, and Mark O Riedl. 2018. Rationalization: A neural machine translation approach to generating natural language explanations. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 81–87.

[28] Andrew J Elliot, Mark D Fairchild, and Anna Franklin. 2015. *Handbook of color psychology.* Cambridge University Press.

[29] Michael W Floyd and David W Aha. 2016. Incorporating transparency during trust-guided behavior adaptation. In *International Conference on Case-Based Reasoning*. Springer, 124–138.

[30] Matthias Galster, Danny Weyns, Dan Tofan, Bartosz Michalik, and Paris Avgeriou. 2014. Variability in software systems—a systematic literature review. *IEEE Transactions on Software Engineering* 40, 3 (2014), 282–306.

[31] Melinda T Gervasio, Karen L Myers, Eric Yeh, and Boone Adkins. 2018. Explanation to Avert Surprise.. In *IUI Workshops*, Vol. 2068. CEUR-WS.org. http://ceur-ws.org/Vol-2068

[32] Alvin I Goldman et al. 2012. Theory of mind. *The Oxford handbook of philosophy of cognitive science* (2012), 402–424.

[33] Ze Gong and Yu Zhang. 2018. Behavior Explanation as Intention Signaling in Human-Robot Teaming. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1005–1011.

[34] Antoine Grea, Laëtitia Matignon, and Samir Aknine. 2018. How explainable plans can make planning faster. In *Workshop on Explainable Artificial Intelligence*. 58.

[35] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)* 51, 5 (2018), 93.

[36] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* (2017).

[37] Maaike Harbers, Jeffrey M Bradshaw, Matthew Johnson, Paul Feltovich, Karel Van Den Bosch, and John-Jules Meyer. 2011. Explanation in human-agent teamwork. In *International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems*. Springer, 21–37.

[38] Maaike Harbers, Karel Van Den Bosch, and John-Jules Meyer. 2009. A methodology for developing self-explaining agents for virtual training. In *International Workshop on Languages, Methodologies and Development Tools for Multi-Agent Systems*. Springer, 168–182.

[39] Maaike Harbers, Karel van den Bosch, and John-Jules Meyer. 2010. Design and evaluation of explainable BDI agents. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. IEEE, IEEE Computer Society Press, 125–132.

[40] Maaike Harbers, Karel van den Bosch, and John-Jules Ch Meyer. 2009. A study into preferred explanations of virtual agent behavior. In *International Workshop on Intelligent Virtual Agents*. Springer, 132–145.

[41] Maaike Harbers, Karel van den Bosch, and John-Jules Ch Meyer. 2011. A Theoretical Framework for Explaining Agent Behavior.. In *SIMULTECH*. SciTePress, 228–231.

[42] Jacob Haspiel, Na Du, Jill Meyerson, Lionel P Robert Jr, Dawn Tilbury, X Jessie Yang, and Anuj K Pradhan. 2018. Explanations and Expectations: Trust Building in Automated Vehicles. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 119–120.

[43] Helen Hastie, Francisco J Chiyah Garcia, David A Robb, Atanas Laskov, and Pedro Patron. 2018. MIRIAM: A Multimodal Interface for Explaining the Reasoning Behind Actions of Remote Autonomous Systems. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 557–558.

[44] 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*. ACM, 303–312.

[45] Steven R Haynes, Mark A Cohen, and Frank E Ritter. 2009. Designs for explaining intelligent agents. *International Journal of Human-Computer Studies* 67, 1 (2009), 90–110.

[46] Aroua Hedhili, Wided Lejouad Chaari, and Khaled Ghédira. 2013. Explanation language syntax for Multi-Agent Systems. In *2013 World Congress on Computer and Information Technology (WCCIT)*. IEEE, 1–6.

[47] Thomas Hellström and Suna Bensch. 2018. Understandable robots. *Paladyn, Journal of Behavioral Robotics* 9, 1 (2018), 110–123.

[48] Koen V Hindriks. 2012. Debugging is explaining. In *International Conference on Principles and Practice of Multi-Agent Systems*. Springer, 31–45.

[49] Daniel Holliday, Stephanie Wilson, and Simone Stumpf. 2013. The effect of explanations on perceived control and behaviors in intelligent systems. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. ACM, 181–186.

[50] Andreas Holzinger, Bernd Malle, Peter Kieseberg, Peter M Roth, Heimo Müller, Robert Reihs, and Kurt Zatloukal. 2017. Towards the Augmented Pathologist: Challenges of Explainable-AI in Digital Pathology. *CoRR* abs/1712.06657 (2017).

[51] Frank Kaptein, Joost Broekens, Koen Hindriks, and Mark Neerincx. 2017. Personalised self-explanation by robots: The role of goals versus beliefs in robot-action explanation for children and adults. In *Robot and Human Interactive Communication (RO-MAN), 2017 26th IEEE International Symposium on*. IEEE, IEEE, 676–682.

[52] Frank Kaptein, Joost Broekens, Koen Hindriks, and Mark Neerincx. 2017. The role of emotion in self-explanations by cognitive agents. In *Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), 2017 Seventh International Conference on*. IEEE, 88–93.

[53] B.A. Kitchenham, P. Brereton, M. Turner, M.K. Niazi, S. Linkman, R. Pretorius, and D. Budgen. 2010. Refining the systematic literature review process-two participant-observer case studies. *Empirical Software Engineering* 15, 6 (2010), 618–653. https://doi.org/10.1007/s10664-010-9134-8

[54] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman. 2009. Systematic literature reviews in software engineering - A systematic

literature review. *Information and Software Technology* 51, 1 (2009), 7–15. https://doi.org/10.1016/j.infsof.2008.09.009

[55] Ross A Knepper. 2016. On the communicative aspect of human-robot joint action. In *the IEEE International Symposium on Robot and Human Interactive Communication Workshop: Toward a Framework for Joint Action, What about Common Ground*.

[56] Ross A Knepper, Christoforos I Mavrogiannis, Julia Proft, and Claire Liang. 2017. Implicit communication in a joint action. In *Proceedings of the 2017 acm/ieee international conference on human-robot interaction*. ACM, 283–292.

[57] Raj Korpan and Susan L Epstein. 2018. Toward Natural Explanations for a Robot's Navigation Plans. (2018).

[58] Jens Kröske, Kevin O'Holleran, and Hannu Rajaniemi. 2009. Trusted Reasoning Engine for Autonomous Systems with an Interactive Demonstrator. In *4th SEAS DTC Technical Conference*. Citeseer.

[59] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *Visual Languages and Human-Centric Computing (VL/HCC), 2013 IEEE Symposium on*. IEEE, 3–10.

[60] Pat Langley, Ben Meadows, Mohan Sridharan, and Dongkyu Choi. 2017. Explainable Agency for Intelligent Autonomous Systems.. In *AAAI*. AAAI Press, 4762–4764.

[61] Sau-lai Lee, Ivy Yee-man Lau, Sara Kiesler, and Chi-Yue Chiu. 2005. Human mental models of humanoid robots. In *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*. IEEE, 2767–2772.

[62] Benjamin Lettl and Axel Schulte. 2013. Self-explanation capability for cognitive agents on-board of UCAVs to improve cooperation in a manned-unmanned fighter team. In *AIAA Infotech@ Aerospace (I@ A) Conference*. 4898.

[63] Sirui Li, Weixing Sun, and Tim Miller. 2015. Communication in human-agent teams for tasks with joint action. In *International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems*. Springer, 224–241.

[64] Christina Lichtenthäler, Tamara Lorenzy, and Alexandra Kirsch. 2012. Influence of legibility on perceived safety in a virtual human-robot path crossing task. In *RO-MAN, 2012 IEEE*. IEEE, 676–681.

[65] Brian Y Lim and Anind K Dey. 2011. Design of an intelligible mobile context-aware application. In *Proceedings of the 13th international conference on human computer interaction with mobile devices and services*. ACM, 157–166.

[66] Zachary C Lipton. 2018. The mythos of model interpretability. *Queue* 16, 3 (2018), 30.

[67] Meghann Lomas, Robert Chevalier, Ernest Vincent Cross II, Robert Christopher Garrett, John Hoare, and Michael Kopack. 2012. Explaining robot actions. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. ACM, 187–188.

[68] Neil A Macmillan. 2002. Signal detection theory. *Stevens' handbook of experimental psychology* 4 (2002), 43–90.

[69] Martin Michlmayr. 2002. Simulation theory versus theory theory: Theories concerning the ability to read minds. (2002).

[70] Masahiko Mikawa, Yuriko Yoshikawa, and Makoto Fujisawa. 2018. Expression of intention by rotational head movements for teleoperated mobile robot. In *Advanced Motion Control (AMC), 2018 IEEE 15th International Workshop on*. IEEE, 249–254.

[71] Tim Miller. 2018. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* (2018).

[72] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of inmates running the asylum, In IJCAI-17 Workshop on Explainable AI (XAI). *CoRR* 36.

[73] Matthew Molineaux, Dustin Dannenhauer, and David W Aha. 2018. Towards Explainable NPCs: A Relational Exploration Learning Agent. (2018).

[74] Mark A Neerincx, Jasper van der Waa, Frank Kaptein, and Jurriaan van Diggelen. 2018. Using Perceptual and Cognitive Explanations for Enhanced Human-Agent Team Performance. In *International Conference on Engineering Psychology and Cognitive Ergonomics*. Springer, 204–214.

[75] Mehrbakhsh Nilashi, Dietmar Jannach, Othman bin Ibrahim, Mohammad Dalvi Esfahani, and Hossein Ahmadi. 2016. Recommendation quality, transparency, and website quality for trust-building in recommendation agents. *Electronic Commerce Research and Applications* 19 (2016), 70–84.

[76] Tatsuya Nomura and Kayoko Kawakami. 2011. Relationships between Robot's Self-Disclosures and Human's Anxiety toward Robots. In *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Volume 03*. IEEE Computer Society, 66–69.

[77] Jekaterina Novikova, Leon Watts, and Tetsunari Inamura. 2015. Emotionally expressive robot behavior improves human-robot collaboration. In *Robot and Human Interactive Communication (RO-MAN), 2015 24th IEEE International Symposium on*. IEEE, 7–12.

[78] Mayada Oudah, Talal Rahwan, Tawna Crandall, and Jacob W Crandall. 2018. How AI Wins Friends and Influences People in Repeated Games with Cheap Talk. In *Proceedings of the 32nd National Conference on Artificial Intelligence*.

[79] Sean Penney, Jonathan Dodge, Claudia Hilderbrand, Andrew Anderson, Logan Simpson, and Margaret Burnett. 2018. Toward Foraging for Understanding of StarCraft Agents: An Empirical Study. In *23rd International Conference on Intelligent User Interfaces*. ACM, 225–237.

[80] David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences* 1, 4 (1978), 515–526.

[81] David V Pynadath, Ning Wang, Ericka Rovira, and Michael J Barnes. 2018. Clustering Behavior to Recognize Subjective Beliefs in Human-Agent Teams. In *Proceedings of the 17th International Conference on Autonomous Agents and Multi-Agent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1495–1503.

[82] Lara Quijano-Sanchez, Christian Sauer, Juan A Recio-Garcia, and Belen Diaz-Agudo. 2017. Make it personal: a social explanation system applied to group recommendations. *Expert Systems with Applications* 76 (2017), 36–48.

[83] Anand S Rao, Michael P Georgeff, et al. 1995. BDI agents: from theory to practice.. In *ICMAS*, Vol. 95. 312–319.

[84] Jorge Rios-Martinez, Anne Spalanzani, and Christian Laugier. 2015. From proxemics theory to socially-aware navigation: A survey. *International Journal of Social Robotics* 7, 2 (2015), 137–153.

[85] Eric Rosen, David Whitney, Elizabeth Phillips, Gary Chien, James Tompkin, George Konidaris, and Stefanie Tellex. 2017. Communicating robot arm motion intent through mixed reality head-mounted displays. *arXiv preprint arXiv:1708.03655* (2017).

[86] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296* abs/1708.08296 (2017).

[87] Rebecca Saxe, Laura E Schulz, and Yuhong V Jiang. 2006. Reading minds versus following rules: Dissociating theory of mind and executive control in the brain. *Social neuroscience* 1, 3-4 (2006), 284–298.

[88] Aroua Hedhili Sbaï, Wided Lejouad Chaari, and Khaled Ghédira. 2013. Intra-agent explanation using temporal and extended causal maps. *Procedia Computer Science* 22 (2013), 241–249.

[89] Rosario Scalise. [n. d.]. Human-Centered Design of Robot Explanations. ([n. d.]).

[90] Raymond Sheh. 2017. Different XAI for different HRI. In *AAAI Fall Symposium-Technical Report*. 114–117.

[91] Raymond Sheh and Isaac Monteath. 2017. Introspectively Assessing Failures through Explainable Artificial Intelligence. In *IROS Workshop on Introspective Methods for Reliable Autonomy*.

[92] Ivan Shindev, Yu Sun, Michael Coovert, Jenny Pavlova, and Tiffany Lee. 2012. Exploration of intention expression for robots. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. ACM, 247–248.

[93] Sichao Song and Seiji Yamada. 2018. Effect of Expressive Lights on Human Perception and Interpretation of Functional Robot. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, LBW629.

[94] Sarath Sreedharan, Tathagata Chakraborti, and Subbarao Kambhampati. 2017. Balancing explicability and explanation in human-aware planning. In *2017 AAAI Fall Symposium*. AI Access Foundation, 61–68.

[95] Simone Stumpf, Weng-Keen Wong, Margaret Burnett, and Todd Kulesza. 2010. Making intelligent systems understandable and controllable by end users. (2010).

[96] Roykrong Sukkerd, Reid Simmons, and David Garlan. 2018. Towards Explainable Multi-Objective Probabilistic Planning. In *Proceedings of the 4th International Workshop on Software Engineering for Smart Cyber-Physical Systems (SEsCPS\ '18)*.

[97] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[98] Leila Takayama, Doug Dooley, and Wendy Ju. 2011. Expressing thought: improving robot readability with animation principles. In *Human-Robot Interaction (HRI), 2011 6th ACM/IEEE International Conference on*. IEEE, 69–76.

[99] Rik Van den Brule, Gijsbert Bijlstra, Ron Dotsch, Daniël HJ Wigboldus, and WFG Haselager. 2013. Signaling robot trustworthiness: Effects of behavioral cues as warnings. *LNCS* 8239 (2013), 583–584.

[100] Jo Vermeulen. 2010. Improving intelligibility and control in ubicomp. In *Proceedings of the 12th ACM international conference adjunct papers on Ubiquitous computing-Adjunct*. ACM, 485–488.

[101] Paul Voigt and Axel Von dem Bussche. 2017. The EU General Data Protection Regulation (GDPR). *A Practical Guide, 1st Ed., Cham: Springer International Publishing* (2017).

[102] 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 2016 international conference on autonomous agents & multiagent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 997–1005.

[103] Handy Wicaksono and Claude Sammut1 Raymond Sheh. [n. d.]. Towards Explainable Tool Creation by a Robot. In *IJCAI-17 Workshop on Explainable AI (XAI)*. 63.

[104] Ryan W Wohleber, Kimberly Stowers, Jessie YC Chen, and Michael Barnes. 2017. Effects of agent transparency and communication framing on human-agent teaming. In *Systems, Man, and Cybernetics (SMC), 2017 IEEE International Conference on*. IEEE, 3427–3432.

[105] Robert H Wortham and Andreas Theodorou. 2017. Robot transparency, trust and utility. *Connection Science* 29, 3 (2017), 242–248.
[106] Robert H Wortham, Andreas Theodorou, and Joanna J Bryson. 2016. What does the robot think? Transparency as a fundamental design requirement for intelligent systems. (2016).
[107] Robert H Wortham, Andreas Theodorou, and Joanna J Bryson. 2017. Improving robot transparency: real-time visualisation of robot AI substantially improves understanding in naive observers. In *Robot and Human Interactive Communication (RO-MAN), 2017 26th IEEE International Symposium on*. IEEE, 1424–1431.
[108] Quan-shi Zhang and Song-Chun Zhu. 2018. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering* 19, 1 (2018), 27–39.