EXPLAINABLE GOAL-DRIVEN AGENTS AND ROBOTS - A COMPREHENSIVE REVIEW

Fatai Sado
Department of Artificial Intelligence
Faculty of Computer Science and Information Technology
University of Malaya
sado.fatai@ieee.org

Chu Kiong Loo *
Department of Artificial Intelligence
Faculty of Computer Science and Information Technology
University of Malaya
ckloo.um@um.edu.my

Wei Shiung Liew
Department of Artificial Intelligence
Faculty of Computer Science and Information Technology
University of Malaya
liew.wei.shiung@siswa.um.edu.my

Matthias Kerzel
Department of Informatics
Knowledge Technology
University of Hamburg
Vogt-Koelln-Strasse 30
20146
matthias.kerzel@informatik.uni-hamburg.de

Stefan Wermter
Department of Informatics
Knowledge Technology
University of Hamburg
Vogt-Koelln-Strasse 30
20146
wermter@informatik.uni-hamburg.de
*

July 23, 2021

ABSTRACT

Recent applications of autonomous agents and robots, e.g., self-driving cars, scenario-based trainers, exploration robots, service robots etc., have brought attention to crucial trust-related challenges associated with the current generation of artificial intelligence (AI) systems. AI systems based on the connectionist deep learning neural network approach lack capabilities of explaining their decisions and actions to others, despite their great successes. Without symbolic interpretation capabilities, they are 'black boxes', which renders their choices or actions opaque, making it difficult to trust them in safety-critical applications. The recent stance on the explainability of AI systems has witnessed several approaches to eXplainable Artificial Intelligence (XAI); however, most of the studies have focused on data-driven XAI systems applied in computational sciences. Studies addressing the increasingly pervasive goal-driven agents and robots are still missing. This paper reviews approaches on explainable goal-driven intelligent agents and robots, focusing on techniques for explaining and

communicating agents' perceptual functions (e.g., senses, vision, etc.) and cognitive reasoning (e.g., beliefs, desires, intention, plans, and goals) with humans in the loop. The review highlights key strategies that emphasize transparency, understandability, and continual learning for explainability. Finally, the paper presents requirements for explainability and suggests a roadmap for the possible realization of effective goal-driven explainable agents and robots.

1 Introduction

1.1 Background/Motivation

Goal-driven agents (GDAs) and robots are autonomous agents capable of interacting independently and effectively with their environment to accomplish some given or self-generated goals [1]. These agents should possess human-like learning capabilities such as perception (e.g., sensory input, user input, etc.) and cognition (e.g., learning, planning, beliefs, etc.). In addition to problem-solving skills, the agents should provide a causal explanation for their decisions and reasoning. GDAs engage in tasks that require activity over time, generate plans or goals, and execute their ideas in the environment, applying both perception and cognition. They can also adapt the intentions or purposes as the need arises and may be required to account for their actions [2, 3], and embodying such capabilities in the context of lifelong developmental learning [4]. For several purposes, GDAs are helpful, including space and mine exploration, agent debugging, scenario-based training, agent development, transportation, and gaming [3]. A relevant example in this context involves an autonomous robot that plans and carries out an exploration task. It then participates in a debriefing session to provide a summary report and addresses a human supervisor's questions. Agents must explain the decisions they made during plan generation, stating considered alternatives, to report which actions they executed and why, explain how actual events diverged from the plan and how they adapted in response, and communicate decisions and reasons in a human-understandable way to gain user's trust [5, 6, 7].

This review focuses on two aspects of human-like learning for goal-driven agents and robots: explainability and continual learning of explanatory knowledge. Explainable AI enables explainability. We focus on both situated/non-situated and embodied/non-embodied autonomous eXplainable GDAs (XGDAIs). The review categorizes explanation generation techniques for XGDAIs according to the agent's behavioural attributes, i.e. reactive, deliberative, and hybrid. It provides a clear taxonomy on XGDAIs and clarifies the notion of the agent's behavioural attributes that influences explainability (Section 2.2). For each category, the review focuses on explanation at the level of the agent's perception (e.g., sensory skills, vision, etc.) and cognition (e.g., plans, goals, beliefs, desires, and intentions). While an agent's perceptual foundation (dominated by reactive reasoning) may be connected to the sub-symbolic reasoning part relating the agent's states, vision, or sensors/environmental information to the agent's cognitive base, the cognitive base (dominated by deliberative reasoning) relates plans, goals, beliefs, or desires to executed actions. Finally, we provide a roadmap recommendation for the effective actualization of explainable autonomous GDAs with an extended perceptual and cognitive explanation capability.

1.2 What is eXplainable AI?

eXplainable AI (XAI) refers to machine-learning or artificial intelligence systems capable of explaining their behavior in human-understandable ways [8]. Explanations help humans collaborating with an autonomous or semi-autonomous agent to understand why the agent failed to achieve a goal or completes a task unexpectedly. For instance, a non-expert human collaborating with an agent during a search-and-rescue mission demands trust and confidence in the agent's action. If the agent fails to finish up the task or unpredictably performs the task, it is natural for the human collaborator to understand why. Explanations thus enable the human collaborator to comprehend the dynamics leading to the agent's actions, enabling the human to decide how to deal with that behavior.

1.3 Why Explainability?

Although the need for explainability of AI systems has been long established during the MYCIN era, also known as the era of the expert systems [9, 10], the current drive for explainability has been motivated by recent governmental efforts from the European Union, United States (USA) [11], and China [12] which have identified artificial intelligence (AI) and robotics as economic priorities. The European Union's key recommendation is the right to explanation which is stressed by the General Data Protection Regulation (GDPR) [13, 3]. AI systems must explain their decisions, actions, or predictions for safety-critical applications to ensure transparency, explainability, and accountability [12].
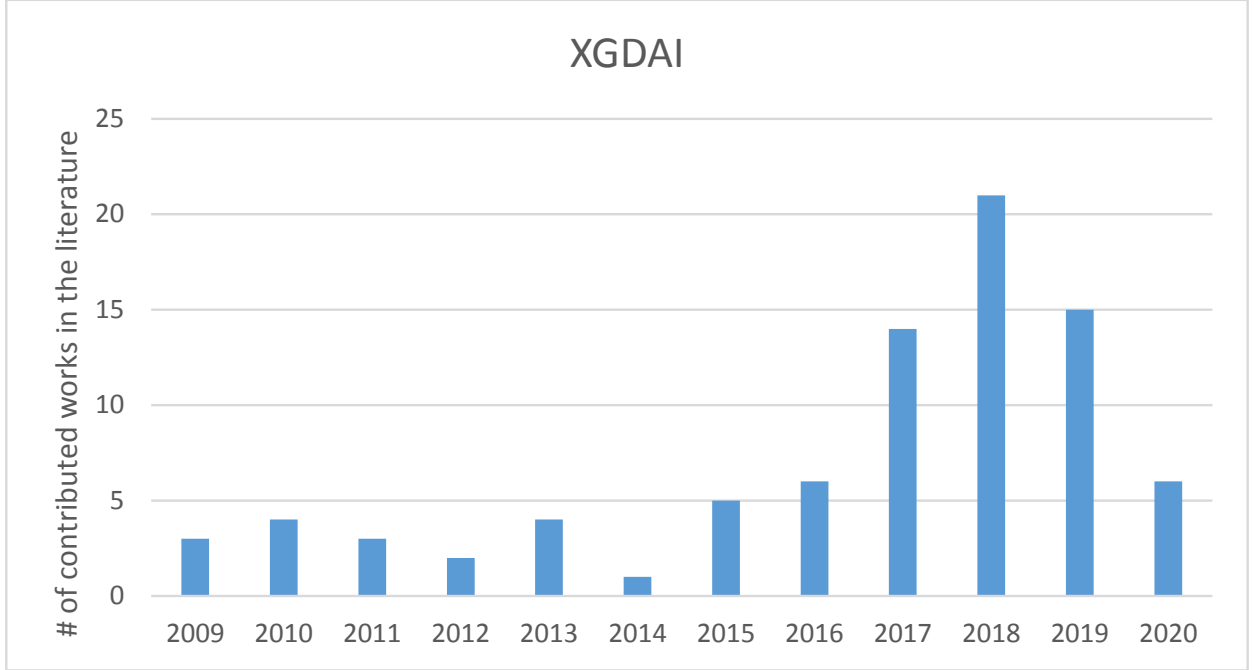
Figure 1: Emerging trend in XGDAI. The chart indicates the progression of the total number of publications over the last decade whose keywords, title, and/or abstracts refer to the XGDAI field. Data collected from Web of Science and Scopus (August 28th, 2020) while querying the databases using the search term indicated by the legend. The increasing number of studies on explainable goal-driven AI over the last five years is important to note.

1.4 What is Data-driven and Goal-driven XAI?

In machine learning, explainability in data-driven AI is often linked to interpretability. According to Biran and Cotton [1], a particular system is interpretable if a human can understand its operations through explanation or introspection. Choo and Liu [5], for instance, described a deep learning neural network's interpretability as determining the input attributes that account for output predictions. Thus, data-driven XAI implies explaining the decision made by a machine learning "black-box" system, given input data [14]. The motivation to find out how available data contributes to a decision is an important aspect of this branch of XAI and whether the machine learning process can reliably replicate the same decision, given similar data and specific circumstances [15].

On the other hand, goal-driven XAI is a research domain that aims to create explainable robots or agents that can justify their behaviors to a lay user [13]. The explanations would assist the user to create a Theory of the Mind (ToM), comprehend the agent's behavior, and contribute to greater collaboration between the user and the agent. A better ToM of the agent would enable the user to understand the agents' limitations and capabilities, thus enhancing confidence and safety levels and preventing failures. The absence of adequate mental models and understanding of the agent can contribute to a failure of interactions [16, 13]. As discussed in the next subsection, there is a growing application of goal-driven XAI in the current AI-dependent world [204].

1.5 Emerging Trends in XGDAI

Fig. 1 shows the chronological distribution of works on eXplainable goal-driven AI (XGDAI) over the last decade. The distribution shows an uneven proportion in the number of studies before 2014. However, there is an increasing number of studies on XGDAI over the last five years. This upsurge in publication can be seen as the effect of the general pressure on the explainability of AI systems and initiatives by several national government agencies like the "right to explanation" by the GDPR [13] [3]. This trend may likely increase in the upcoming years, with several researchers working on XAI in different research domains.

2 Terminology Classification

2.1 Terms in XGDAI

Different terminologies can be found in the literature for the description of Goal-driven AI. Some authors use terms such as goal-directed agents or robots [17, 18, 19, 20], goal-seeking agents [21, 22, 23], or simply autonomous agents [24]. An important attribute of these agents is that they seek to achieve a goal or execute a plan. With respect to explainability in XGDAI, terms such as understandability [25], explicability [26], transparency [27], predictability [19], readability [28], and legibility [19] can be found. These terminologies are used interchangeably with explainability. In this section, the distinction and similarities between these terms are clarified.

Explainability

Explainability can be understood as an agent's capability of making decisions that are understandable to humans. The explanation acts as an interface between the artificial decision-making system and humans [22]. The explanation provided can take different forms - e.g. linguistic (verbal), visual, textual, symbolic gesture, expressive motion, eye gaze, etc. Details of different forms of explanation communication are presented in Section 4.

Understandability

Understandability denotes the notion of communicating an agent's function – how it works - in the way that a human can understand it. This may involve communicating its computational process or internal structure in a human-understandable way [29, 25]. Understandability may also be achieved in an agnostic approach to represent the agent's function regardless of the internal structure or algorithm used to process its data. For example, explanations can be provided as complementary information in a recommendation system to help users better understand what the system is doing and to encourage confidence and trust in the system's assessment [30].

Transparency

Transparency implies the ability to relate or replicate the processes by which decisions are made by an AI agent/robot or the processes by which it learns to interact with the environment [29].

Explicability

According to Sreedharan et al. [31], an explicable system avoids the necessity to offer explanations by constructing plans which are consistent with users' expected plan.

Predictability

According to Dragan et al. [19], predictability refers to the quality of an agent's/robot's behavior or action matching expectations, implying that an agent's/robot's step towards a goal is predictable if it matches what an observer would expect. The term, however, is similar to the notion of explicability.

Legibility

Legibility refers to the quality of a robot behavior or action to be intent-expressive, i.e., the practice can enable an observer to infer its intention. A legible robot motion allows the observer to easily and appropriately infer the robot's goals [19].

Readability

Readability in XGDAI implies the notion that robot behavior is human-readable in such a way that people can figure out what the robot is doing and can reasonably predict the next robot action to interact effectively with the robot. [28].

The overall use of terminologies for XGDAI suggests a trend towards (1) explicit explainability where the agent/robot provides a clear explanation for its behavior – decisions or actions – and (2) implicit explainability where the agent/robot avoids the need to provide explanations by making its behavior readable, legible, predictable, explicable, or transparent.

2.2 Attributes of XGDAI

Several attributes can be found in the literature for an XGDAI. Regarding the agent's behavior and interaction with the world, three behavioral architectures are traditionally distinguished: deliberative – in which the agent deliberates (plans ahead to reach its goals) on its goals, plan, or action, or acts based on a sense-model-plan-act cycle (the agent, in this case, should possess a symbolic representation of the world); reactive – In which the agent implements some simple behavioral patterns and reacts to activities or events in its environment in a stimulus-response way, no model of the world is required (the robot chooses one action at a time); and hybrid - which combines the above two behaviors [24]. Some other terminologies include goal-driven autonomy, goal-driven agency, and BDI. Table 1 in the supplementary materials presents the taxonomies of XGDAI behavior found in the literature. In this section, we make further clarification of these attributes.

Reactive

Reactive agents present a collection of simple behavioral patterns that react to environmental changes. [24], no model of the world is included. They can only achieve their goal by reacting reflexively to external stimuli, choosing one action at a time. The creation of purely reactive agents came at the heels of the limitations of symbolic AI. Developers of reactive agent architecture rejected symbolic representations and manipulation as a base of AI [32]. Model-free (deep) reinforcement learning (RL) is a state-of-the-art approaches that enables reactive agent behavior. Some notable explainability works in RL include Memory-based eXplainable Reinforcement Learning (MXRL) [33], Minimal Sufficient eXplanation (MSX) via Reward Decomposition [34], and Reward Augmentation and Repair through Explanation (RARE) [35]. Some previous reviews on reactive RL agents can be found in Lin [36] and Lin [37].

Deliberative

Deliberative agents act more like they think, searching through a behavioral space, keeping an internal state, and predicting the consequences of an action. Wooldridge [38] describes such an agent as one with a symbolic model of its world, and that could make decisions based on symbolic reasoning or rationale. The cognitive aspect of these agents consists primarily of two parts, according to the conventional approach: a planner and a world model [24]. The world's model is mainly an internal representation of, and often involves, the agent itself in the agent's external environment. The planner uses this representation to make a strategy for how the agent's goals can be accomplished. How such agents operate may be viewed as a 'sense-model-plan-act' behavior. The Belief-Desire-Intention (BDI) model is the most widely used architecture to enforce such actions, where the beliefs of an agent about the world (its picture of the world), desires (goal) and intentions are internally expressed, and realistic reasoning is applied to determine which action to select [39].

Hybrid

A significant number of researches has been focused on combining both reactive and deliberative agent techniques, leading to the creation of a compound called a hybrid agent, which combines comprehensive internal manipulation with non-trivial symbolic constructs and external events with reflexive reactive responses [24]. This integration of flexibility and robustness of reactivity and the foresight of deliberation is suggested to be the modern drive [40] to integrate the flexibility and robustness of reactivity with the foresight of deliberation. Supporters of the hybrid approach believe it is optimal since high-level deliberative reasoning and low-level response to perceptual stimuli seem essential for actualizing an agent's behavior. An example is the hybrid system proposed by Wang et al. [23]. Reactive exploration is used to create waypoints, which are then used by a deliberative method to plan future movements in the same area. Another existing system with mixed reactive and deliberative behaviors is the agent developed by Rao et al. [41], which figures out when to be reactive and when to pursue goal-driven plans.

2.3 Application Scenarios for XGDAI

This section presents the application scenarios for XGDAI that are primarily reported in the literature. XGDAI application scenarios include robot-human collaborative tasks, search and rescue, E-health, gaming applications, training, ubiquitous computing, recommender systems, and robot navigation. These are summarized in Table 1 in the supplementary materials.

Robot-human collaborative tasks

Tasks involving close interactions with humans in factory settings and teaming in an outdoor setting are the predominantly mentioned application scenarios in literature. In robot-human collaborative scenarios, explainability (i.e.,

transparency) of XGDAI was shown to enhance the quality of task performance [112] and to enable both robots or humans to take responsibility (credit or blame) for their actions in collaborative situations [113].

Robot navigation

In robot navigation, Korpan and Epstein [40] propose a "Why-Plan" that contrasts the viewpoints of an autonomous robot and a human while co-planning a navigation route and explains why the plan deviates. A goal-discovering robotic architecture is proposed in Santucci et al. [85] to explore the world autonomously and learn various skills that make it possible for the robot to change the world. A neural network controller for detecting unexpected events during robot navigation is proposed in Jauffret et al. [84], enabling the agent to solve complicated navigation tasks and seek assistance if it encounters situations of deadlock.

Game application

In a game application, non-player characters' actions are explained to reduce player misunderstanding [87].

Search and Rescue

In a search and rescue simulation environment, an Explaining Robot Action (ERA) system is implemented to enable robots to address questions concerning their actions [62].

Training

In a virtual training system, natural language explanations are used as intention signaling to understand robot behavior [46].

Ubiquitous computing

In ubicomp systems, it is suggested that intelligibility helps people to comprehend the system's operation and allows users to intervene when the system makes an error. [103].

E-health

In e-health, a PAL (Personal Assistant for a Safe Lifestyle) agent that communicates with children, their parents, and their caregivers to assist them with the treatment process is proposed, along with an explanation that acknowledges the robot's own and others' emotions [51].

Recommender systems

An explanation facility for a recommender system called Personalized Social Individual Explanation (PSIE) is proposed for group recommendation [30]. For movie and music recommender systems, an explanation is introduced in [108] to investigate how users' mental models were impacted by varying completeness and soundness.

3 Explanation Generation Techniques for XGDAI

This section presents existing explanation generation techniques and taxonomies (e.g., transparency, domain dependence, post-hoc explainability, continual learning, etc.) for XGDAI. The section is further subdivided into two parts. The first part discusses current deliberative XGDAI techniques, and the second subsection discusses reactive XGDAI techniques. The overview of explainability techniques for XGDAI shows that the techniques are domain-specific, agnostic, or post-hoc. Domain-specific explainability techniques heavily depend on the agent world's domain knowledge and do not permit application to other agents in other environments. Domain agnostic or post-hoc explainability techniques are domain-independent, allowing cross-platform explainability. Post-hoc explanations make it possible to explain without inherently following the reasoning process leading to the decision [13].

3.1 Deliberative XGDAI

3.1.1 Models of Goal-driven Autonomy - Transparent domain-agnostic

Goal-driven autonomy is defined as a conceptual goal justification mechanism that enables an agent to generate explanations when detecting a mismatch in the environment state. The model enables the robot to continuously track

the execution of its current plan, determine whether the states met expectations, and generate an explanation upon identifying a mismatch [88]. The model is also extended to enable an agent to identify when new goals should be chosen and explain why new goals should be followed [64].
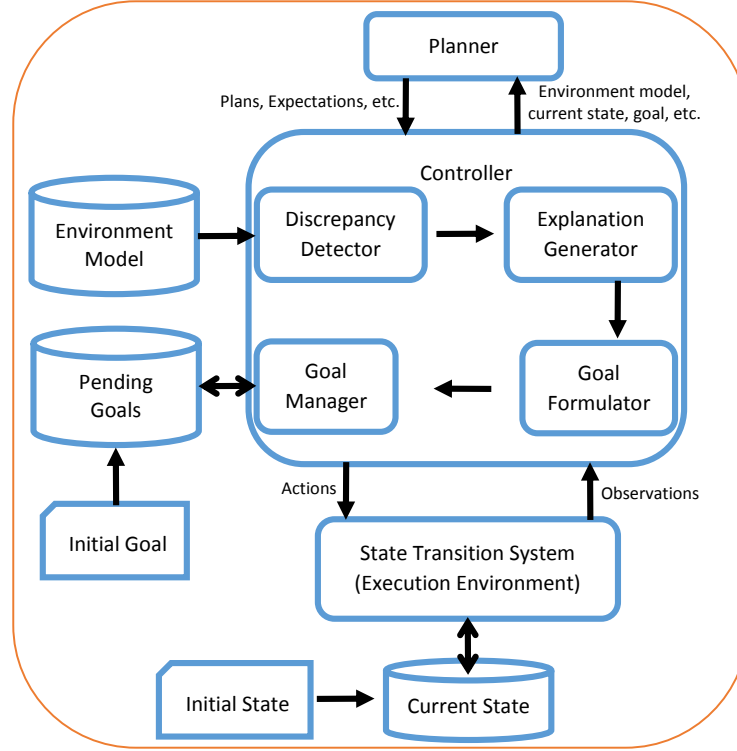

Figure 2: Goal-driven autonomy conceptual model [88]. The model extends the Nau [116] online planning model. The controller communicates with the execution environment (i.e., state transition system) and the planner. As input, the planner receives a planning problem consisting of an environmental model, the current state of the system, and a potential goal. The planner generates a series of action plans and expectations (states constraints when executing actions). The controller sends the action plan to the state transition system, then processes the subsequent observations. The controller detects discrepancies, produces explanations, generates goals and manages the goals.

In many existing works in goal-driven autonomy, explanations are generated when active goals are completed or when discrepancies are detected between the actual event and the intended plan. Such differences may result if the domain knowledge is flawed, i.e., if the dynamics governing the states projection is flawed or the perception of a state is incorrect. Discrepancies may also result if there is a hidden factor influencing the state. An explanation is generated to explain the discrepancies and find or address the hidden factors that affect the state [88]. The Autonomous Response to Unexpected Events (ARTUE) [88] is a domain-independent goal-driven agent that continuously reacts to unexpected situations in the environment by reasoning on what goals to pursue to resolve the situation. ARTUE is designed to deal with unexpected environmental situations by explaining these changes first and then developing new goals which integrate the explicit information on the hidden environmental aspects. On this basis, ARTUE should deal with new and unobserved objects in the planning. The ARTUE system is implemented in a Sandbox simulation scenario [114]. The system (Fig. 2) integrates a planner that deliberates on exogenous events through predicting future states in a dynamic environment and a controller that identifies discrepancies, generates explanation, generates targets and manages goals. The explanation aspect deliberates on the world's hidden knowledge by abductive reasoning on the conditions and effects of the planning [115].

3.1.2 Explainable BDI model - Transparent model-specific

BDI agents or robots, primarily symbolic AIs, with integrated beliefs, desires, and intentions offer good prospects for producing useful and human-understandable explanations [94]. According to Bratman et al. [117], belief and desire are both mental attitudes (pro-attitudes) that drive an action. Intention is distinguished as a conduct controlling this pro-attitude, which can be treated as elements of partial plans of action [117, 118]. Harbers et al. [94] proposes a BDI model that allows the explanation of a BDI agent's actions (behavior) based on the underlying beliefs and goals. Goals

are described as "active" desires that the agent currently pursues. An example of an explanation of actions based on 'belief' can be: "The boy opened the window because he believed someone was outside", and based on 'goals' can be: "The boy opened the window because he wanted to see who was outside". The motivation here is that humans can clarify and comprehend their behaviors or actions in terms of underlying beliefs, desires, intentions, or goals [119, 120]. Thus, since BDI agents establish their actions by deliberation on their mental state, the mental principles behind the action are applied to interpret their actions. Also, since mental reasoning is expressed symbolically, the explanation generation process is therefore straightforward. Typically, a log of behaviour preserves all the agent's prior mental states and actions that could be used for explanations. On request, the BDI algorithm is implemented on the log and selects the beliefs and goals that become part of the explanation. However, not all 'explanatory elements' can be helpful in the explanation [119].

An important aspect of explainable BDI agents is that they can clarify typical human errors. According to Flin and Arbuthnot [121], revealing explainable BDI agents' actual mental state to trainees may make them aware of their (false) assumptions about them. People may make false assumptions about others' experience and motives in many crucial circumstances [121], a phenomenon of attributing incorrect mental states to others [122].

3.1.3 Situation Awareness–based Agent Transparency (SAT) model - Transparent model-specific

SAT is an agent transparency model for enhancing an operator's situational awareness [123] in the agent's environment [60, 59]. SAT model depends on the BDI model framework and implemented as a user interface to provide transparency to the operator. An important extension is that it provides transparency not only of the status of the robot (e.g., plans, current state, and goals) and process of reasoning, but also of the future projections of the robot (e.g., future environment states) [47]. Basic information on the robot's current goal and state, its intentions and expected action are presented to the operator at the entry phase of the SAT structure. At the second stage, knowledge about the agent's reasoning process supporting its action and the limitations it takes into consideration are presented before the operator. At the third stage, specifics of the robot's future projection are given to the operator, such as expected outcomes, the possibility of failure or success, and any ambiguity inherent in the predictions. The transparency of an agent in this context is an informative ability to provide an operator's understanding of the purpose, results, plans, reasoning process, and future projections of an agent. While the operator's trust is described as the agent's willingness to help achieve the operator's goals in especially uncertain and vulnerable circumstances [13].

3.1.4 Meta-AQUA Explanation model - Transparent domain-agnostic

Meta-AQUA is an introspective learning system proposed by Cox [124] for a self-aware cognitive agent. The system can allow an agent to decide (learn) its goals by understanding and describing unusual occurrences in the world. To minimize the dissonance between what the agent expects and the world's actual reality, the learned objectives seek to modify the agent's perception of the world. With intelligent behaviors, Meta-AQUA integrates cognitive components such as planning, understanding, and learning and metacognition (such as cognition monitoring and control). Learning in this context is described as a deliberate planning task with a set of learning goals, whereas the explanation of unusual events is a key to enhance the agent's learning of a new goal. Meta-AQUA uses its metareasoning component (i.e. metacognition) to explain a reasoning or expectation failure, enhancing its goal learning. It adopts case-based knowledge representations made as frames linked together by explanatory patterns to represent general causal structures [125, 124].

Meta-AQUA is implemented in the INitial inTROspective cognitive agent (INTRO) and simulated in the Wumpus World-simulated environment, a partially observable environment where the objective of the agent or robot is to seek a pot of gold while evading pits and the Wumpus creature (Fig. 3). INTRO is designed with primitive perceptual and effector subsystems and two cognitive components.

There are four main components of INTRO: a primitive subsystem of perceptions and effectors and two components of cognitive planning and comprehension. The cognitive components compose of the respective Prodigy/Agent and Meta-AQUA structures. The Perceptual subsystem act as a translator for the cognitive subsystems. The central cognitive cycle is to explore the world, form an objective of changing the world by understanding and describing unusual events or world states, establish a plan to accomplish the goal, and finally perform according to the plan, in turn observing the outcomes. Instead of using explanation to alter its understanding, INTRO uses explanation to create a goal to improve the environment.

The agent will observe the world with these components, develop a goal to transform the world, devise a strategy for achieving the objective, and eventually change the environment, including turning, going forward, picking up, and firing an arrow. As the agent navigates the environment, the Meta-AQUA system seeks to explain the effects of the agent's behavior by inserting the events into a conceptual representation and constructs an internal model to reflect the causal
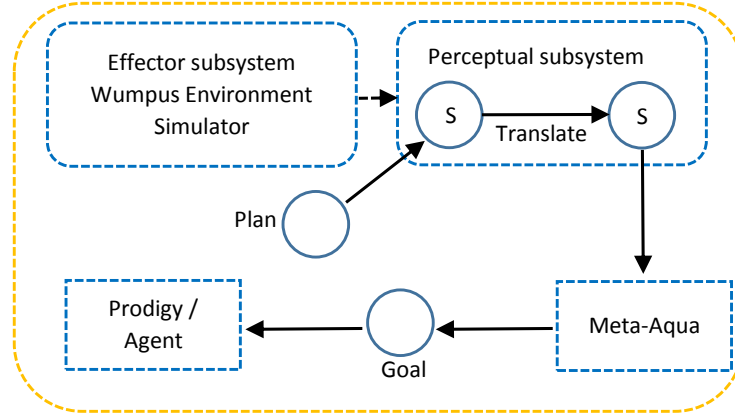
Figure 3: The INTRO architecture [124] consists of a primitive subsystem of perceptions and effectors and two components of cognitive planning and comprehension.

ties between them. Meta-AQUA creates an anomaly or other interesting occurrence to produce an explanation of the event. However, a major aspect of this approach is that the agent utilizes the explanation to produce a goal to modify the environment (e.g. shoot an arrow, etc.) instead of using the explanation to change its understanding.

3.1.5 Proactive Explanation model

In scenarios that involve teams of humans and autonomous agents, a proactive explanation that anticipates potential surprises can be useful. By providing timely explanations that prevent surprise, autonomous agents could avoid perceived faulty behavior and other trust-related issues to enable effective collaboration with team members. This is an important extension to explainability for agents since most techniques are usually provided as a reaction to users' queries. In this context, Gervasio et al. [67] presents a conceptual framework that proposes explanations to avert surprise, given a potential expectation violation. Proactive explanation here attempts to reconcile the human mental model with the agent's real decision-making process to minimize surprise that could disrupt the team's collaboration. Surprise is used as the primary motivation for pro-activity and is triggered if the agent's action deviates from its past actions, or if its action would be considered unusual or contrary to plan [67].

3.1.6 KAGR Explanation model – Post-hoc domain-agnostic

KAGR is an explanation structure proposed by Sbai et al. [110] to explain agent reasoning in multi-agent systems (MAS) (operating in uncontrollable, dynamic, or complex situations) where an agent's reasoning is not reproducible for the users. KAGR describes an agent's reasoning state at runtime as a quadruplet $< K, A, G, R >$ (Knowledge, Action, Goal, Relation), also considered explanatory knowledge. Under the KAGR conceptual explanation framework, events produced by the agents at run-time are intercepted, and an explanatory knowledge acquisition phase is performed to represent the knowledge attributes or details related to the execution of events in a KAGR structure [111, 110]. A further step is performed to generate a knowledge representation formalism by linking the attributes in an extended causal map (CM) model. A final step achieves natural language interpretation for the CM model using predicate first-order logic to build up a knowledge-based system for an understandable reasoning explanation to users. Overall, as seen in Fig. 4, it adopts a three-module architecture: a first module that generates explanation, another module that puts the knowledge in the formalism of the extended causal chart, and a third module that uses first-order logic to analyze and interpret the constructed causal maps to produce reasonable explanations.

3.1.7 eXplainable Plan models – Post-hoc/Transparent domain-agnostic

An aspect of planning is Plan Explanation, in which the primary objective is to enable individuals to understand the plans of the (agent) planner (e.g., [126], [127], [128]), including interpreting the agents' plans into human-understandable form and the creation of interfaces that facilitate this understanding. In this context, related works include XAI-PLAN [48], RBP [128], and WHY-PLAN [40].

XAI-PLAN is an explanatory plan model proposed by Borgo et al. [48] to provide an immediate explanation for the decisions taken by the agent planner [48]. The model produces explanations by encouraging users to try different alternatives in the plans and after that compares the subsequent plans with those of the planner. The interactions between
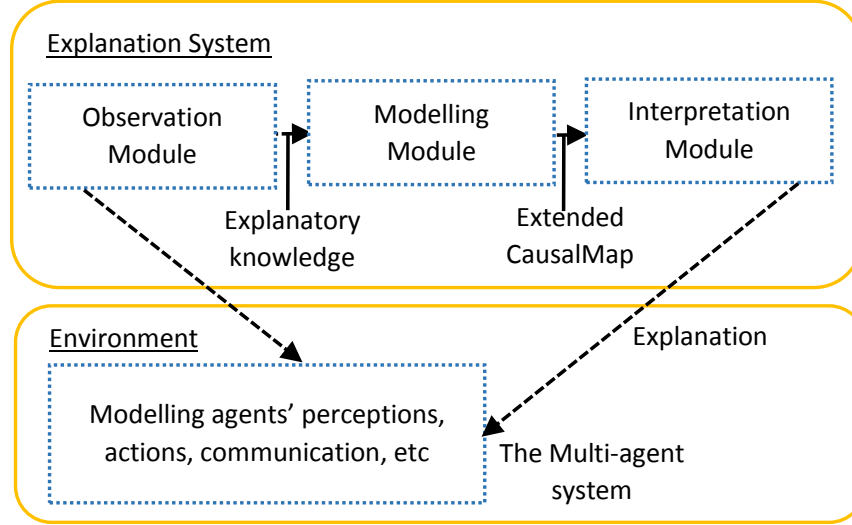
Figure 4: KAGR explanation architecture [110]. The system consists of an observation module for the generation of the explanatory knowledge, a modelling module for the knowledge formalism using extended causal map, and an interpretation module to produce explanation.
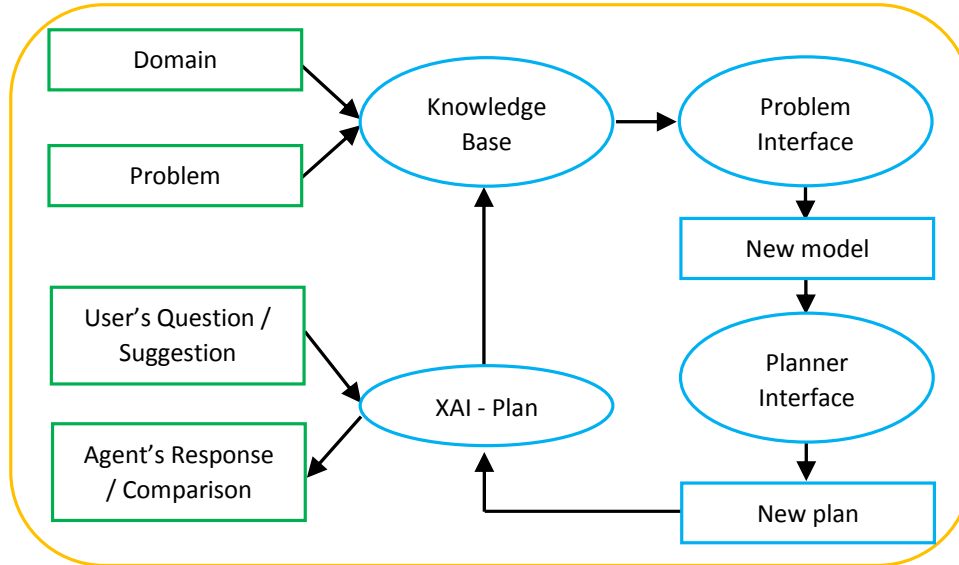

Figure 5: Example XAI – Plan architecture by Borgo et al. [48]. The XAI-Plan node generates explanatory plans and communicates agent's response through a user interface. ROSPlan provides the planner interface, problem interface, and knowledge base.

the planner and the user enhance hybrid-initiative planning that can improve the final plan. The XAI-PLAN planning system is domain-agnostic and independent. XAI-PLAN provide answers to questions such as "why does the plan includes a specific action and not another similar action?". The algorithm uses a preliminary set of plans as input; the user chooses one of the plan's actions; the XAI-Plan node generates explanatory plans and communicates with the user through the user interface (Fig. 5). ROSPlan provides the planner interface, problem interface, and knowledge base, which is used to store a Planning Domain Definition Language (PDDL) model and to provide the AI planner (i.e., an architecture for embedding task planning into ROS systems) with an interface.

Refinement-based planning (RBP) is a transparent domain-independent framework proposed by Bidot et al. [128] to allow verbal queries from users and produce plan explanations verbally. RBP enables a transparent description of the

search space examined during the planning process, giving the possibility to explore the search space backward to search for the relevant flaws and plan modifications. RBP builds on a hybrid planning structure incorporating hierarchical planning and planning using partial-order causal-link using states and action primitives [129].

Korpan and Epstein [40] propose WHY-Plan as an explanation method that contrasts the viewpoints of an autonomous robot and a human while planning a navigation route. The core of its explanation is how the planners' objectives differ. WHY-Plan addresses the question "Why does your plan go this way?" and exploits differences between planning objectives to produce meaningful, human-friendly explanations in natural language. An example of a WHY-Plan natural language explanation for a robot, in a team with a person, navigating from one location to another and avoiding a crowded path could be, "Although a somewhat shorter route exists, I think mine is a lot less crowded". The WHY-Plan's response compares two objectives: "avoid crowds", which is its planning objective and an alternative objective: "take the shortest path", which can be attributed to the human questioner or team member. WHY-Plan is implemented in SemaFORR, a cognitive-based hybrid robot controller [130]. SemaFORR could make a decision using two sets of reactive advisors.

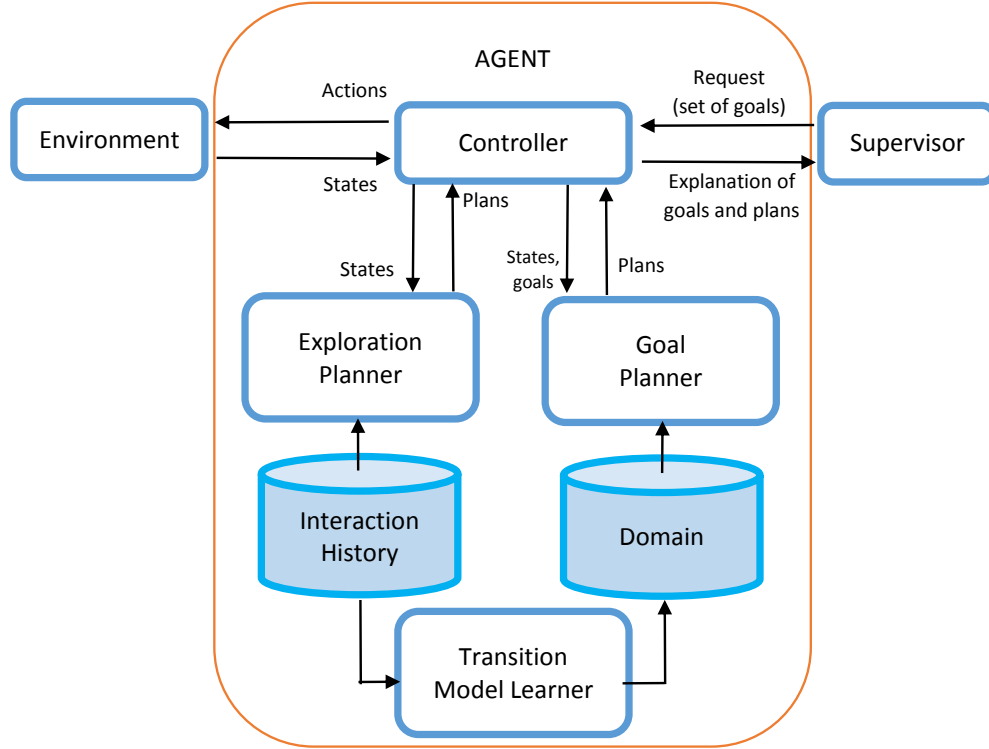### 3.1.8 Explainable NPC – domain-independent/agnostic

Figure 6: Example Explainable NPC Architecture [87]. The agent has four sub-modeled functionalities: the exploratory planner for obtaining new information; the goal planner necessary for implementing the supervisor's goals; the transition model learner responsible for updating the world model of the agent; and the controller responsible for determining when to explore

Explainable Non-Player Characters (Explainable NPC) is an architecture proposed by Molineaux et al. [87] to minimize the frustration of video game players by explaining NPC actions. To many video game players, non-player characters (NPCs) may become a source of frustration because their reasoning process is not transparent. The NPCs may respond to some internal necessity that a player is unaware of or may face barriers that a player can not see, but these problems are not communicated by them [87]. The NPC architecture is thus motivated to allow agents to know their environment, achieve a set of goals and demonstrate to a supervisor what they are doing (Fig. 6). At each step, the architecture enables the agent to obtain an observation that represents knowledge about the environment's true state and to communicate with the environment (and supervisor) by taking action. The supervisor requests the agent that represents what the supervisor wants the agent to achieve, modified at each step. In response, the agent should always clarify why it takes a specific action to the supervisor. The framework is divided into four submodules: the exploratory planner, tasked with taking steps to collect new data to update an action model, the goal-directed planner, necessary for implementing the

supervisor's goals, the learner of the transition model, necessary for updating the world model of the agent, and the controller, responsible for deciding when to explore [87].

In related work, Van Lent et al. [131] propose the eXplainable AI (XAI) architecture for NPCs in a training system. The Explainable AI (XAI) functions to retrieve important events and decisions made from the replay log (in the post-action assessment) and enable NPCs to explain their conduct as an answer to the questions selected from the XAI menu. In this framework, each character's reactive behaviour, low-level actions, and higher-level behavior generation are the responsibility of the non-player character (NPC) AI. During the execution process, the XAI system records the NPC AI's activities and utilizes the records or logs for post-action analyses.

3.1.9   Debrief – Transparent domain-specific

Debrief is a multimedia explanation system proposed by Johnson [97] that constructs explanations after an active goal by recalling the circumstance in which the decision was taken and replaying the decision in the variants of the original situation or by experimenting to decide which factors were crucial to the decision. The elements are considered critical as their absence would result in a different outcome of the decision process. The details of the agent's implementation, including individual rules and motivations applied in making a decision, are filtered out to construct an explanation. Under the framework, it is unnecessary to maintain a complete trace of rule firings to produce explanations. The system learns the relationships between situational factors and decisions to facilitate the explanation of similar decisions. Debrief is implemented in an artificial pilot simulator where the pilot can perform a flight patrol mission. After each task, the pilot is asked to describe and explain essential decisions it made along the way. The pilot has to clarify its evaluation of the situation. Debrief should enable the pilot to define the reasons for its actions and assess the situation and its beliefs that resulted in the actions.

3.1.10   Explainable CREATIVE model – Transparent domain-specific

Cognitive Robot Equipped with Autonomous Tool InVention Expertise (CREATIVE) is a relational approach proposed by Wicaksono and Sheh [50] enabling robots to learn how to use tools and design and construct new tools if needed. The critical information, such as learned observations, snapshots from a camera, and objects positions, are stored in Prolog to simplify the explanation of its tool creation process. CREATIVE utilizes relational representation, so its results have some inherent explainability features, establishing the relation between entities as facts in Prolog (Programming in Logic). A method of inductive logic programming (ILP) is employed to learn the relational representations of the tool models [132]. A pre-determined set of basic questions and answers is used in the entire dialogue in Prolog.

3.1.11   Summary on Deliberate XGDAIs

Overall, reviewed literature on deliberative XGDAIs provide applicable explanation generation models and frameworks focusing on transparency and/or comprehensibility of agents' plans, goals, actions (behavior), reasoning process, decisions, current state, and world state (i.e. unusual events and future projections of states). Leveraging on their model for plan execution (or model of the world or tasks), their primary triggers or key basis for explanation generation include factors such as expectation failures (i.e. mismatches between what the agent expects and what the agent encountered), unusual events (in the world that is not previously modeled for the agent), anticipated surprise for agent's action or decision, mismatched mental models (of the agent and human), users' requests, and underlying belief or goal driving an action. These techniques assume certain aspects of the human-like explanation generation process. However, most of the approaches are mostly domain-dependent, over-fitted to the particular situation or environment, limiting application to other domains.

Another concern for deliberative XGDAIs is how explanatory knowledge is acquired for explanation generation. Handcrafting explanatory knowledge or storing logs of agents' internal states (e.g. mental states and actions) or prolog of facts appears the predominant technique. In this context, resource management for storing explanatory knowledge is a significant concern that is yet to be addressed in the literature. A summary of key findings on deliberative XGDAI is presented in Table 2 in the supplementary materials.

3.2   Reactive XGDAI

3.2.1   Automated Rationale Generation model - Post-hoc domain-agnostic

Automated rationale generation (ARG) is an explainable model proposed by Ehsan et al. [8] for real-time explanation generation, which is trained to interpret the inner state and data representation of an agent in natural language. The approach is domain agnostic and ideally developed for a 'Markovian' sequential environment where agent's states and actions are accessible to the model and where decisions (i.e. selection of actions that maximize expected future
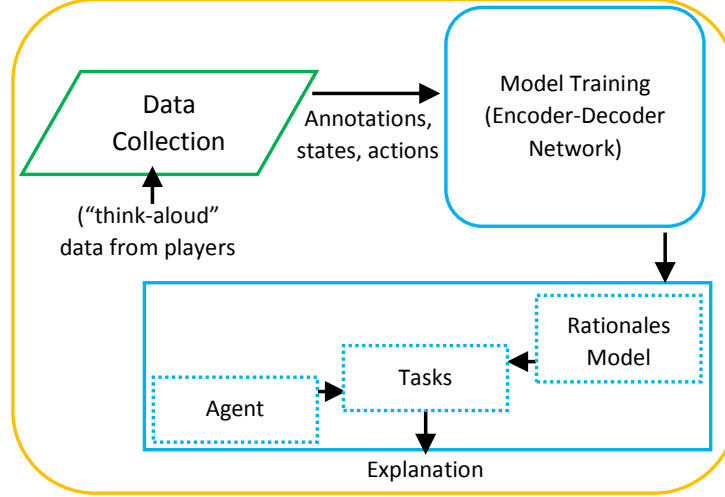
Figure 7: Automated rationale generation architecture of Ehsan et al. [8]. An end-to-end training method for explanation generation. The technique is to gather a think-aloud data corpus (i.e. actions, game states, and annotations) of players that discussed their actions during the game and train the encoder-decoder network to enable agents to produce reasonable rationales for their actions.

rewards or utility) made by the agent in the past have an impact on future decisions. ARG is expected to generate an explanation or rationale in the natural language for the behavior of agents as if a person had performed such behavior [92]. The idea behind the rationale generation is that people can communicate effectively by verbalizing reasonable motives for their actions, even though the verbalized reasoning is not aligned with the decision-making neural processes of the human brain [134]. Ehsan et al. [8] applied ARG for human-like explanation generation for an agent playing Frogger, a sequential game environment. The ARG approach is to collect a corpus of human-like explanation data (Fig. 7) - i.e. A thought-aloud data corpus of players that discussed their behavior during the game and train a neural rationale generator (an encoder-decoder neural network) using the corpus to enable the agents to produce their plausible human-like rationales or explanations.

While the results are promising, the the rationale generator's potential limitations may be the lack of a more grounded and technical explanation framework. The framework lacks interactivity that offers users the possibility to question a rationale or to suggest that an agent clarifies its decision in another way. Another limitation may stem from the data collection process that introduces breakpoints to request an explanation from the players. It is necessary to determine an appropriate method to collect data from players during runtime without interruption of the players or participants.

3.2.2 Explainable Reinforcement Learning (XRL) – Post-hoc domain-independent

Explainable reinforcement learning (XRL) is proposed in several studies for a class of reactive reinforcement learning (RL) agents [33]. Reactive RL agents are model-free agents that determine what to do depending on their current observations [135]. They rely on a simple behavioral policy scheme (i.e., a state-to-action mapping) that allows them to learn a policy based on trial-and-error interaction with their environment [136]. RL agents do not generally reason or plan for their future actions, which makes it challenging to explain their behavior. An RL agent may eventually learn that one action is preferred over others or that choosing an action is associated with a higher value to attain the goal but would lose the rationale behind such a decision at the end of the process [136]. Thus, they lack the mechanism that can effectively explain why, given a specific state, they choose certain actions. Some existing techniques of explainability for RL agents aim to make the decision process during the policy learning process retrievable and explainable. Some examples include MXRL [33], Minimal Sufficient Explanation (MSX) via Reward Decomposition [34], and RARE [35].

Memory-based eXplainable Reinforcement Learning (MXRL) is an explainable reinforcement learning (XRL) strategy proposed by Cruz et al. [33] to enable a reinforcement learning agent to explain to users why it selected an action over other possible actions. The model uses an episodic memory [136] to save each episode or agent's record of executed state-action combinations, then compute both the likelihood of success (Q-values) as well as the number of transitions within each episode to meet the final target to provide an explanation or reason for selecting an action over the others. MXRL is implemented in two discrete simulated environments: a bounded grid world and an unbounded grid

environment with aversive regions. The RL agent can clarify its actions to lay users at any time by using information obtained from memory. MXRL, however, suffers from limitations with regard to the utilization of memory in broad solution spaces.

Minimal Sufficient Explanation (MSX) is an XRL strategy proposed by Juozapaitis et al. [34] to explain RL decisions agents via reward decomposition. The idea is to break down incentives into amounts of semantically meaningful reward types, allowing actions to be contrasted in terms of trade-offs between them. In a domain-independent setting, MSX is supposed to provide a concise description of why one behaviour is favoured over another in terms of reward styles. It makes use of an off-policy version of Q-learning that leads to the best policy and decomposed action values. The focus is on explanations that learn Q-functions that allow observing the actions preferences of the agent quantitatively. MSX is implemented in two environments to support its validity: a CliffWorld grid-world where cells can contain cliffs, monsters, gold bars, and treasure that is decomposed into reward types [cliff, gold, monster, treasure] reflecting the current cell's contents; and a Lunar Lander rocket scenario where the actions can be decomposed into natural reward types including crashing penalty, safe landing bonus, main-engine fuel cost, side-engines fuel cost, and shaping reward that defines scenarios close to the actual world) of controlling a rocket during a ground landing.

Reward Augmentation and Repair through Explanation (RARE) is an extension of the Explainable Reinforcement Learning (XRL) strategy [35] to address the need for establishing a shared behavior (mental) model between an RL agent and a human collaborator (for effective collaboration) using a timely update to their reward function. The RARE framework is modeled on the premise that sub-optimal collaboration activity results from a misinformed understanding of the task/assignment rather than a problem with the reasoning of the collaborator. Thus, using the Markov decision-making process, sub-optimal human decision-making is attributable to a malformed policy due to an inaccurate task model. RARE will determine, through interactive learning and reward updates, the most probable reward function for human actions, identify the missing aspects of the reward function, and convey it as actionable information to allow the collaborator to update their reward function (task understanding) and policy (behavior) while performing the task and not after the task is completed. This mechanism should allow the agent or robot to provide a policy update to a person (i.e. by explaining the correct reward to the human) based on perceived model difference, minimizing the risk of costly or dangerous errors during the execution of joint tasks. A color-based collaborative version of Sudoku and an autonomous robot are used to implement RARE. (Rethink Robotics Sawyer) [35]. The robot is reported to interrupt users who are on the verge of making a mistake, inform them that their actions will cause a failure of the task, and explain which constraint of the game will inevitably be violated. However, the RARE model still lacks the comprehensibility of its optimal policies. During the computation of an optimal policy, the factors taken into account for risk and reward considerations (for each state and prospective action) are lost.

Other XRL strategies for model-free RL agents include the work of Madumal et al. [137] that utilizes causal models to generate 'contrastive' explanations (e.g., "why" and "why not" questions) as a means of describing partly measurable agent action or actions in a game scenario (Starcraft II). The approach is to learn a structural causal model (SCM) during reinforcement learning and to generate explanations for "why" and "why not" questions by counterfactual analysis of the learned SCM. However, one weakness of the approach is that the causal model must be given beforehand. In another work by Pocius et al. [138], deep RL were used to provide visual interpretations in the form of saliency maps to describe an agent's decisions in a partially observable game scenario. However, saliency maps do not help to explain long-term causality and can be sensitive. The study by Sequeira and Gervasio [136] provided explanations through introspective analysis of the RL agent's interaction history. The framework explores the history of an agent's environmental experiences to retrieve interesting elements that explain its behaviour.

3.2.3 Autonomous Policy Explanation – Post-hoc domain-agnostic

Autonomous Policy Explanation is a strategy proposed by Hayes and Shah [74] for a class of (reactive agent) robot controllers that rely on black-box trained reinforcement learning models [21], or on hard-coded conditional statement-driven policies, to enable the robot to autonomously synthesize policy descriptions and respond to natural language queries by human collaborators. The objective is for the robot to reason over and answer questions about its underlying control logic independently of its internal representation, allowing human coworkers to synchronise their perceptions (or mental models) and recognise defective actions in the robot controller. The model applies to discrete, continuous, and complex dynamic environments. Using a Markov Decision Process (MDP) model for constructing the domain and policy models of the control software, the system learns a domain model (i.e., collection of states) of its operating environment and the robot's underlying control logic or policy from actual or simulated demonstrations or observations of the controller's execution traces. These are incorporated into a single graphical model that collects the essential details about states' relationships and behaviour.

For natural language communication, the model employs communicable predicates, i.e., Boolean classifiers similar to traditional STRIPS-style [139] planning predicates with associated natural language descriptions, to translate attributes
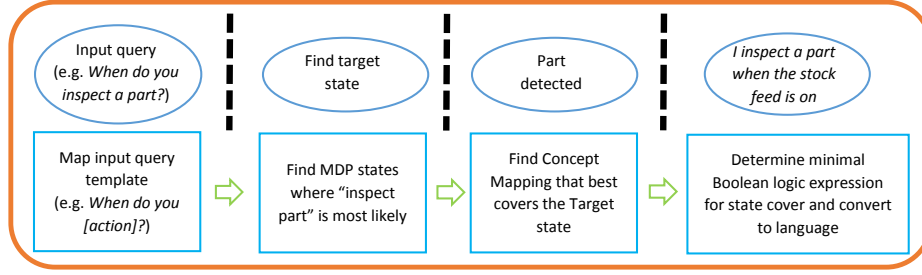
Figure 8: Automated policy generation framework of [74]. The framework maps a policy clarification to an action/input request. An input request or query (e.g. "When do you 'inspect a part'?") is first matched against pre-defined query templates (e.g. "When do you [action]?"). An algorithm for graphical search is used to find the target state, i.e. the state regions that fulfill the query criteria. Using logical combinations of communicable predicates, a concept mapping that best covers the target state is found. The cover is then reduced and translated via the template into language.

from sets of states to natural language and back (Fig. 8). The algorithms can then enable the agent to answer a direct inquiry about behavior-related questions, e.g. "When do you do _?" - requesting an explanation about the occurrence of a specific controller behavior, or "Why didn't you do _?" - requesting an explanation for why a certain controller activity was not observed.

3.2.4 Summary on Reactive XGDAIs

Overall, reviewed literature on reactive XGDAIs relies on simple behavioural policy models to generate explanation. Most approaches are agnostic and have the potentials to be applied in several domains. However, several gaps are still visible. For example, some models lack a more grounded and technical explanation framework [8] (e.g. an interactive explanation platform, breakpoint-free explanatory knowledge collection process, etc.), some others, particularly RL agents, cannot articulate rationales for their actions or 'concerns' (i.e. comprehensibility of agent's optimal policy after policy convergence). A summary of key findings on reactive XGDAIs can be seen in Table 3 in the supplementary materials.

3.3 Hybrid XGDAI

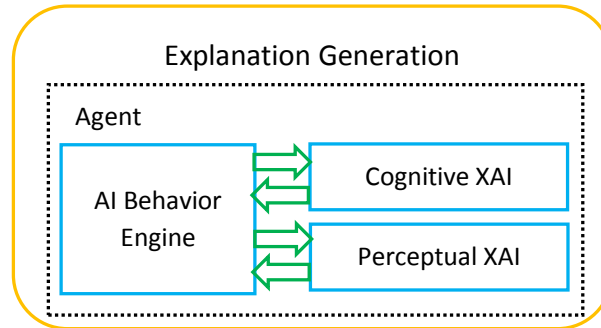### 3.3.1 Perceptual-cognitive explanation (PeCoX) – Domain-agnostic

Figure 9: PeCoX explanation generation framework[82]. The framework distinguishes between Perceptual eXplainable AI (XAI) and Cognitive eXplainable AI (XAI). The Perceptual XAI is connected to the sub-symbolic reasoning of the AI behavior engine and is designed to clarify the perceptual aspect of the agent's behaviour. The Cognitive XAI interacts with this part to ground its belief base. It can explain why certain actions have been selected by linking them to goals or beliefs.

PeCoX is a framework proposed by Neerincx et al. [82] for the creation of explanations that explore the cognitive and perceptual dimensions of the behaviour of an agent (Fig. 9). PeCoX's perceptual XAI explains the perceptual aspect of the agent's behavior using a proposed Intuitive Confidence Measure (ICM) [140] and a method of contrastive explanations [141] that involve the classification of facts and foils, e.g. 'How did you arrive at this outcome (the fact)

rather than the other (the foil)?'. PeCoX's perceptual XAI is model-agnostic, relying entirely on any learned model's input, output, and prospective feedback on that output. The ICM (or uncertainty) measured the expected performance of the machine learning model on any specific classification or decision.

PeCoX's cognitive XAI selects goals, beliefs, and 'emotions' to explain why the agent chooses a certain action. This is similar to the BDI explanation models [94]). The model considers explanations from the intentional stance [142], i.e. the notion that an agent's action depends on its built-in intention, beliefs, or goals. PeCoX XAI includes another cognitive function: 'emotion', e.g. 'I expect (emotion) you would exercise regularly since I want (goal) you to be physically fit and I suspect (belief) you are not currently exercising adequately' [143, 144]. PeCoX's cognitive framework is designed to be domain-agnostic.

4 Explanation Communication Techniques for XGDAI

This section discusses what type of explanation is communicated by agents to users and how it is communicated. This is described by the explanatory form (e.g., visual, textual, speech, etc.) and the explanatory content (e.g., logs, text, etc.). This section distinguishes explanation communication for XGDAI into visual, verbal, gestural, numerical, and textual expressions. A summary of our findings is presented in Table 4 in the supplementary materials.

4.1 Visualization

Visualization is a technique of communicating an agent's plan, "mind", or decision-making system by externalizing the decision support process's pathways. This technique builds visual mediums between the agent and the humans to establish trust and transparency between the humans and the machine. An important justification for plan visualization, for instance, is the need to minimize the time taken to communicate the agents' plans in natural language to the humans in the loop. In some domains, visual explanations are well suited to be integrated into the typical workflow of experts: for example, in the medical domain, experts are used to analyse the results of imaging procedures (Xray, fMRI, etc.), drawing their attention to potentially relevant parts of the image can be an intuitive interface.

Visualization techniques in this section are further divided into neural network approaches and symbolic approaches. Neural network approaches provide low-level explainability of a model's internal representation using visual cues, and otherwise, leave interpretation up to the user. In symbolic approaches, knowledge is encapsulated into a user-interpretable format, i.e., symbols directly computed using traditional connectionist models.

4.1.1 Neural Network Approaches

Class Activation Maps

Deep learning architectures based on convolutional neural networks (CNN) were able to achieve optimal results in many computer vision applications. From the training data, CNNs learn to extract a deep hierarchy of task-relevant visual features [188, 189, 190, 191]. While these feature-extracting filters can be visualized, they are hard to interpret: in lower layers, the filters are mostly edge detectors, while in higher layers, they are sensitive to complex features. Moreover, the filters represent what image features the CNN is sensitive to, but not what features lead to a given image classification. Class Activation Maps (CAM) [148] address this issue by creating a heatmap of discriminative image regions over the input image that shows what parts of the input image contributed how much to the CNN's classification of the image to belong to a selected class. In this way, CAMs supply a visual explanation for a classification. Furthermore, by calculating and comparing Class Activation Maps for different classes, relevant portions of the input image can be visualized for distinction.

CAMs work by inserting a global average pooling (GAP) layer directly after the architecture's final convolutional layer. This layer computes a spatial average of all filters from the previous layer, which is then weighted by the selected output class. A drawback of this approach is that the neural network architecture is altered, and models need to be retrained. Gradient-weighted Class Activation Mappings (Grad-CAM) was introduced by Selvaraju et al. [147] based on gradients flowing into the final convolutional layer to solve this problem. They extend their approach by fusing it with Guided Backpropagation (Guided Grad-CAM) to enhance the resolution of the approach further. An example application field for CAM and related methods is medical image analysis. Ng et al. [146] use a three-dimensional CNN to analyze MRI data of possible migraine patients. They use CAMs to highlight discriminative brain areas. In such applications, CAMs guide a medical expert's attention to different image regions to assess the network's prediction.

Attention Weights

Recurrent neural networks (RNNs) are a popular deep learning framework for temporal modeling, particularly in medical applications for modeling disease progression [193] and predicting diagnoses based on patient records [165]. One solution to improve interpretability or visualization was to introduce an intermediate attention layer [166, 167]. The attention mechanism decomposes a complex input into a series of accumulated attention weights. For example, the Timeline model [168] aggregates contextual information of patient medical codes using an attention mechanism to predict future diagnoses of future medical visits. Medical practitioners were able to gain insight and anticipate future diagnoses by observing attentional weights of medical codes over time. Theoretical scenarios can be simulated by manipulating attention weights and observing the subsequent prediction changes using attention-weight visualization tools [169, 170].

SHapley Additive exPlanation

Lundberg and Lee [171] utilize SHapley Additive exPlanation (SHAP) values, an extension of Shapley values, as a united strategy to explain any machine learning model output. The SHAP values give the cumulative marginal contributions of individual inputs. Feature importance can be evaluated at the global level by taking the combined SHAP scores for a particular attribute or interaction with other features using a dependence scatter plot.

4.1.2 Neural-Symbolic Approaches

Representations in neural networks are distributed among many neurons, making it difficult to isolate and interpret without relying on post-hoc knowledge extraction methods. Symbolic models such as graphs and decision trees are easier to interpret, as each node or symbol represents a single distinct concept. Neural-symbolic (NS) models combine the efficient learning and inference power of traditional neural networks and the interpretability of symbolic knowledge extraction and reasoning [192, 188].

Visual Grounding via Neural-symbolic Structured Graphs

Visual grounding is the general task of locating a structured description's components in an image [151]. The structured description or query can be a natural-language phrase that has been parsed as a symbolic structured graph such as a scene graph or action graph. In human-robot interaction scenarios, visual grounding can enable a robot to accurately locate an object amongst other objects in the workspace [150]. Scene graphs were first used for the visual-grounding challenge by Johnson et al. [152]. The scene graph captures the visual scene's intricate semantics by expressly modeling objects, their descriptive properties, and their interconnections in the scene. For example, a scene described by the sentence "man riding black horse" is converted to a scene-graph representation with nodes corresponding to objects ("man"), descriptions or attributes ("horse is black"), and interconnections or relationships ("man riding horse"). The grounding task creates bounding boxes corresponding to the specified objects, such that the located objects have the specified attributes and relationships.

A neural-symbolic "end-to-end" approach to visual grounding was proposed by Xu et al. [153] to generate structured scene representations from input images. The model addresses scene graph inference by using conventional RNNs that learn to generate image-grounded scene graphs and iteratively improve their predictions via message passing. A Region Proposal Network (RPN) [154] is used to generate a collection of object proposals from an image, transfers to the graph-inference module the extracted features of object regions, and then outputs a graph of the scene consisting of categories of objects, semantic relations between object-pairs, and their bounding boxes. A significant aspect of the work is the messages passing of contextual information between two sub-scene graphs (bipartite graphs) and how RNNs can refine their predictions. In the same context, Yang et al. [155] presented Graph R-CNN, a model for constructing scene graph that can also perform object detection and infer object relations in images. The main framework consists of a Relation Proposal Network (RePN) dealing with the quadratic number of possible relationships between the objects and an attentional Graph Convolutional Network (aGCN), which extracts context information between objects and their relations. For instructional videos, Zhou et al. [156] proposed a weakly-supervised framework for reference-aware visual grounding, where the only thing supervised is the temporal synchronisation between the video segment and the transcription. The framework consists of a visually grounded action graph, i.e. an optimised reference-aware multiple instance learning (RA-MIL) objective for poor control of grounding in videos. A standardised representation capturing the latent dependence between grounding variables and references in videos. Scene graph or action graph representations hold tremendous promise in visual grounding tasks, particularly for enhancing an agent's perceptual-cognitive framework. However, extracting scene graphs from images is still a significant challenge.

Automata

Finite-state machines or automata (FSAs) are computational models that assume one of any number of finite states at any one time and transition to other states in response to particular inputs. Architectures of FSAs are inherently interpretable and illustrated by a series of states linked by transition functions. A large body of work exists for extracting interpretable FSAs from RNNs. For example, using quantization algorithms [157], clustering outputs to infer FSAs [158] [159], and recursively testing inputs and outputs [160].

Transducers are FSAs with both an input and an output, where traditional FSAs have only an input. Transducers can be further divided into Moore machines and Mealy machines. The former produce outputs based solely on the machine state, while the latter relies on both inputs and states. Preference Moore machines [162] can operate as either a simple recurrent neural network or as a symbolic transducer while integrating different neural and symbolic knowledge sources. In a text-mining application [163], symbolic transducers were able to encode concepts from a news corpus quickly. At the same time, neural machines were able to produce improved classification performance with additional training. Transducer extraction [161] was then able to obtain detailed symbolic information while preserving sequential information.

Rules Extraction Algorithms

Rule extraction algorithms are a method for converting the neural network's internal arrangement of neurons into a linguistically-interpretable format. The simplest form of a rule can be described using a conditional "IF-THEN" statement, represented in a neural network (NN) by a single neuron as "IF input=X, THEN output=Y." Andrews et al. [164] classified rule extraction algorithms into three categories. Rule extraction algorithms that work at the neuron-level instead of the whole NN level or global are called decompositional methods. If the neural network is "black-box" (i.e. activation or weights of the network can not be observed directly), then the algorithms are pedagogical. The third category, eclectics, is a combination of decompositional and pedagogical.

4.2 Textual and Verbal Communication

Logs

For agents that rely on relational (or symbolic) logics for knowledge representation and reasoning (KRR), using logs of data comes easily to communicate the internal state of the robots. An example is the explainable BDI model proposed by Harbers et al. [94], where all past mental states and actions of the agent needed for explanations were stored in a behavior log. When requesting an explanation, the algorithm selects beliefs and goals from the logs for the explanation. Other examples are the eXplainable AI (XAI) architecture proposed by Van Lent et al. [131] for NPCs in Full Spectrum Command, which Works to extract main events and decisions made from the replay log during the post-review process.

Prolog

Prolog (Programming in Logic) is a general-purpose logic programming language. An important aspect of this language is the utilization of symbolic representations - e.g. "?-dog(jane). no" corresponding to "Is jane a dog? No - a cat". The Prolog program uses rules and facts and queries that make Prolog search through its facts and rules to work out answers [194]. Wicaksono and Sheh [50] use Prolog for the explainable CREATIVE model to express and store relevant information needed for explanations, such as acquired hypotheses, primitive object poses, and camera snapshots. CREATIVE represents a set of simple questions and answers and describes the relationship between objects as Prolog facts to enable explanation.

Speech

Speech or verbalization has been one of the earliest means of communicating agents' thoughts, beliefs, or actions, especially in the domain of social or service robots. Since social robots must converse with humans daily, they require skills for natural conversational ability. Speech offers a more intuitive and faster means of explanation communication (compared to text) for robots, particularly interacting with the visually impaired or with children who have not yet developed a full reading competency. The earliest generation of humanoid robots (i.e., developed in the 1970s and 1980s) in social robots was equipped with conversational skills, although the skills were usually primitive. They were typically designed as simplistic input/output mapping speech combinations [172].

A good example is the Waseda Robot, WABOT-1 [173], which was designed with the capability to recognize spoken sentences as concatenated words, make vocal responses, and change a related state using a Speech Input-Output System (SPIO) [174]. The WABOT-1 system could accept Japanese spoken command sentences, only in the form of sequences

of individually-spoken Japanese words, and then respond to the meaning of the command verbally to make the robot move as commanded. The system's inner core works as an automaton, making output and transitions after recognizing an input sentence. A further upgrade on the conversational system to make the speech more natural, particularly the speech synthesis part, was introduced in WABOT-2 [175], a robot musician, which could produce speech response by retrieving a word dictionary corresponding to a code of the spoken command. The dictionary stores the names of cv syllabic units (i.e., cv – consonant and vowel) necessary for each unit's words, vowel durations, and accent patterns.

Another more recent social robot in this class is PaPeRo [176], a childcare robot designed to converse with humans more naturally. PaPeRo uses a dictionary of commonly used words and phrases and could also be updated by the designers. Humans interacting with the robot must also converse in similar words and phrases that the robot understands to enable natural conversation. PaPeRo uses an electronic hardware auditory system for human-robot communication [177] and can recognize multiple utterances, can give a quiz to children who provide answers to the quiz using a special microphone, and can tell in natural language the names of the children who got the correct answer. Other similar robots are Honda Asimo [178] which uses a commercial hardware electronic system for speech synthesis, and ASKA [179] receptionist robots, with a conversational speech dialogue system that can recognize users' question and answer the users' question by a text-to-speech voice processing with other additional intuitive channels such as a hand gesture, head movement, and body posture. Robovie [180] is another example in this category that can communicate in English using a vocabulary of about 300 sentences for speaking and 50 words for recognition. The reader is referred to Leite et al. [195] for a more comprehensive review of social robots.

Referring Expression

Referring expression is a language construction or expression used to identify particular objects within a scene [181]. Generation of referring expressions (GRE) or Referring Expression Generation (REG) is expression selection for identifying an element from a set of possible elements when given a shared context between a speaker and the hearer [186]. REG algorithms have been traditionally applied as computational models of people's abilities to refer to objects [196]. A host of scholarly works on REG can be found spanning the last five decades [197]. In recent years, referring expression generation (REG) has received significant attention in human-robot interaction (HRI) scenarios, particularly for situated human-robot dialogues. In situated dialogue, humans and robots/agents are co-present in a shared physical environment or workspace (e.g. in assembly tasks, etc.) and often need to refer to an object in the environment that is being shared or manipulated [182]. The motivation for REG is thus the need for the robot to comprehend expressions concerning objects (in a shared space) and their relationships from inputs of image and natural language. In a contribution by Giuliani et al. [183], a situated reference generation framework is presented for an HRI system to enable it to collaborate with humans in an assembling task that involves building objects from a set of wooden toys. The REG system includes a goal inference sub-symbolic system that can identify people's goals or mistakes by observing their verbal and non-verbal actions. The robot's REG module then uses situated references [184] to clarify to users the mistakes and suggest strategies for solutions.

Another relevant example is the INGRESS (Interactive visual grounding of referring expressions) framework proposed by Shridhar et al. [150]. INGRESS is a robot that executes instructions to pick and place items (in the environment) in natural language. The system can ground referring expressions, i.e. recognising dialogues or references to objects and their relationship from images and natural language inputs, and may also ask questions for interactive clarification of unclear referential expressions. It uses a two-stage long short-term memory (LSTM) framework to establish visual representations of objects and link them to input language expressions to identify the candidate referred-to-objects.

Some other relevant contributions are the gestural-deictic reference by Kranstedt and Wachsmuth [185] and the haptic-ostensive reference by Foster et al. [186], in which the authors argue that a multimodal reference involving manipulation actions with the objects would provide a richer referring approach for conversational partners in situated dialogue. They suggest that a fully-elaborated linguistic reference may not always be necessary between conversational partners.

4.3 Non-verbal cues

Lights

This is an approach to explanation communication that explores the use of lights to visualize the robot's internal state in relation to task and environment. The technique offers a wide variety of options, including color, animation pattern, and speed. A useful justification for expressive light is communicating in the public domain where verbal communication or on-screen display would be helpless (due to robot proximity to humans) to convey the robot state. An example is a robot calling for help. Baraka et al. [53] explored the use of lights as a medium of communication on a mobile robot called CoBot (Fig. 10). The robot communicates with people in various ways, including asking users for assistance, supplying users with valuable information, or suggesting changes to their motions. A node (node 1) gathers the robot's state at
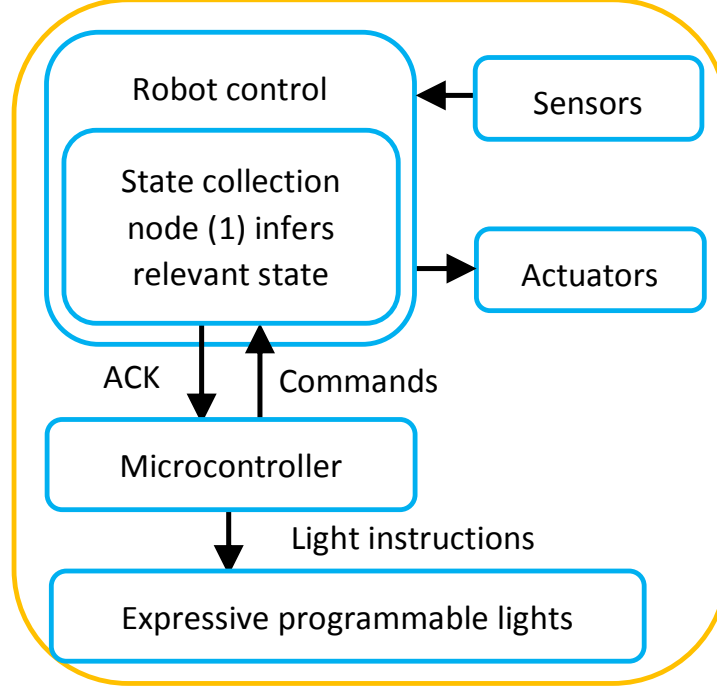
19

Figure 10: CoBot's Control diagram of proposed expressive lights interface [53].

each time phase to communicate its internal state. Whenever a state change occurs, this change triggers a command to the microcontroller (node 2) to inform the controller of the variable(s) that has changed in the state. The microcontroller is programmed with a state-animation mapping algorithm that activates the programmable lights' (3) animation upon notification of a state change.

In related research, Song and Yamada [73] explored the use of expressive light on a Roomba robot (an appearance-constrained robot) to study people's perception and understanding about the robot. Using two light expressions, red and green (for high and low intensity respectively), people could attribute nuanced interpretations of the robot's behavior as the medium of communication. However, the design of expressive lights heavily biases such interpretations.

Expressive motion

One of the concerns for robots wandering among pedestrians is the difficulty in predicting their actions, for example, anticipating which direction the robot will take or what the robot will do. People's ability to understand such robots is also particularly influenced by the robot's appearance. Thus, the motivation here is to make the robot behave more like humans who often employ non-verbal expressions such as changes in facial and bodily expressions. In this context, expressive motion is a useful technique to communicate robots' intentions to pedestrians to improve their mental impressions of the robot. For example, Mikawa et al. [78] proposed using rotational head movement for a teleoperated mobile robot as a way of expressing its intent to pedestrians in public spaces (Fig. 11). The robot communicates its intention by turning its head to see where it moves when traveling around pedestrians. A human operator teleoperates the robot for safety, setting a target position and the robot's movement speed and direction.

Consequently, an artificial potential field (APF) determines the robot head's angle, generated based on the destination and the locations of all obstructions and people around the robot. In effect, people around the robot can discern the robot's intended change of direction in advance. Other related work on expressive motion using gestures and eye gaze can be found in the review of Admoni and Scassellati [187].

5 Continual Learning for Explainability

In this section, we examine techniques in XGDAI that enable continual learning of explanatory knowledge, domain knowledge, domain models, or policies (e.g., sets of environment states, etc.) for explanation generation. This section explores the solution to (1) handcrafting of domain knowledge artifacts for explanation generation in deliberative
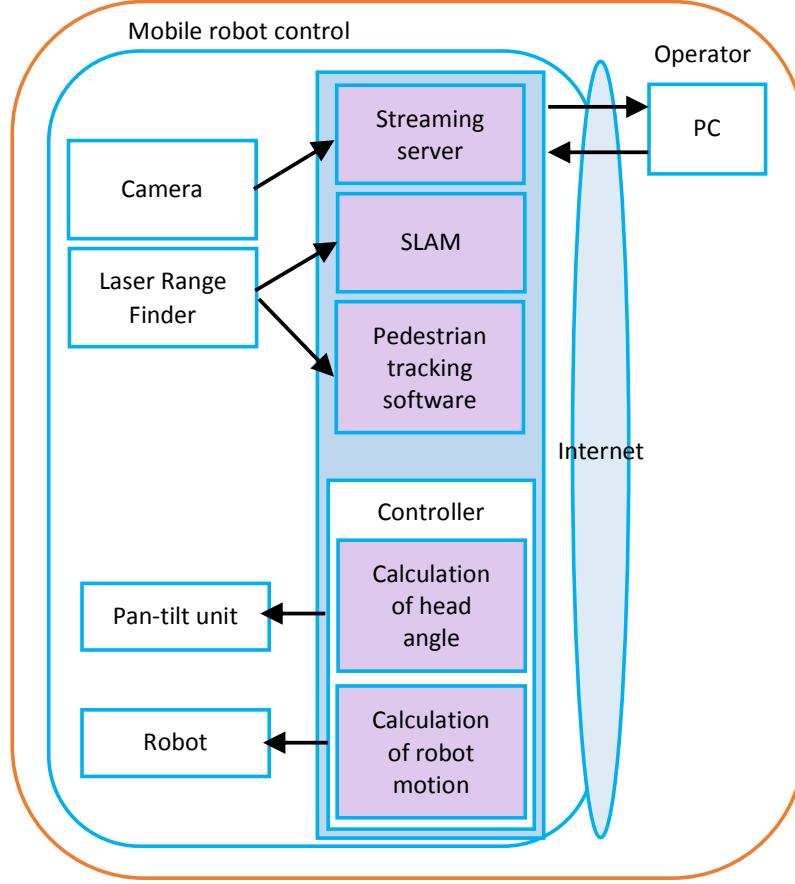
Figure 11: Configuration of mobile robot teleoperation system for expressive head motion [78].

symbolic agents and (2) the solution to learned policy losses during the decision-making process for explainability in reactive RL agents.

5.1 Case-Based Reasoning (CBR)

For agents that rely on handcrafted domain knowledge for defining the explanation components (e.g., expectations, discrepancy definitions, knowledge of how to resolve the discrepancy, operator feedbacks, etc.), the common challenge is that substantial domain engineering is done on the system which needs to be updated each time the robot changes to a new environment. To minimize the prerequisite domain engineering for generating explanations, CBR is used in many XGDAI approaches to acquire domain knowledge at run-time. CBR applies a learning and adaptation strategy to help explanation generation by storing, recalling and adapting knowledge information or experiences (or cases) stored in a case-library or memory [198].

In the CBR framework, agent learning to expand its knowledge is executed by evaluating and integrating new experiences into the case-library and/or re-indexing and reusing previous experiences [199]. CBR has two primary learning methods: observational learning (supervised) and learning through personal experience [198]. Observational learning is performed by filling the case-library with observations from expert demonstrations or from actual data [200]. Learning from one's personal experience [201] happens after a reasoning process that evaluates a potential solution to a challenge. If the solution succeeds, it is saved and applied for future references [198].

Weber et al. [133] proposed an observational learning approach to reduce the number of domain knowledge acquisition necessary for implementing the real-time strategy game StarCraft. The CBR was applied to learn expectations, explanations, and goals from expert demonstrations [133]. An adversary library offers examples of adversary behaviour, and a goal library chooses goals to be pursued by the agent. The learning system offers explanations to the agent if it achieves an intended goal or if a deviation is observed. In a similar effort, Floyd and Aha [76] applied two case-based reasoning (CBR) systems to assist agent learning and explanation generation. Both CBRs use cases that are learned
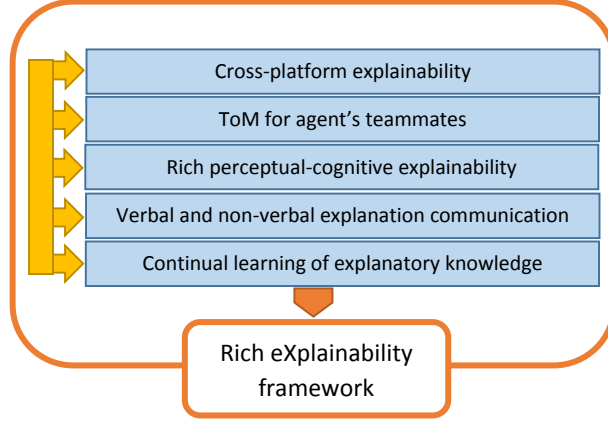
Figure 12: Explainability Framework for XGDAI.

while interacting with the operator (learning by observation). The first CBR process determines if a robot is trustworthy and selects a new behaviour if found otherwise. The second CBR is used to provide an explanation or clarification should the robot change its behaviour. The agent's explanations are based on explicit feedback received from an operator. The model is evaluated in a simulation environment involving an operator who instructs the robot to patrol, detect suspicious objects, and designate the objects as either harmless or dangerous.

5.2 Explainable Reinforcement Learning (XRL)

One of the major concerns for many RL agents is a loss of information about the decision process during the agent's policy learning process. An RL agent may know at the end of a learning objective that certain actions may produce a higher gain or value to attain the goal or that one action is preferred over another but loses the rationale behind the decision-making as the policy converges towards an optimal mechanism for the selection of actions [136]. The lack of bookkeeping, traceability, or recovery of this process, once an optimal policy has been learned, makes it hard for the agent to explain itself or transfer a learned policy for explainability. A few extensions to applying the XRL technique seek to enhance retention of the learned policy for explaining agents' decision processes. Some examples include the Memory-based eXplainable Reinforcement Learning (MXRL) proposed by Cruz et al. [33] that introduced an episodic memory to store important events during the decision-making process of the robot. The episodic memory is designed to enable the agent to introspect, observe, or analyse its environment transitions and interactions. The agent is shown to clarify its actions to lay users during task execution, relying on its episodic memory. The major shortcoming to the MXRL technique, as with many other memory-based continual learning techniques, is the limitations with regards to the use of memory in large solution spaces. There is still an open-ended quest for XRL techniques that enable comprehensibility, continual learning, and policy retention.

6 Discussion

Existing studies in XGDAI show a lack of consensus in the requirements for explainability. Different behavioral architectures for GDAI - e.g. deliberative, reactive, and hybrid - come with different explanation generation techniques. The state-of-the-art suggests the need for an effective unified approach towards explainability in XGDAI. Many explainability techniques still lack an extensive framework: a rich perceptual-cognitive explainable framework, verbal and non-verbal communication framework, framework for natural language processing, and continual learning for explanation construction. In this section, we outline a framework, illustrated in Fig. 12, for the effective actualization of explainability in XGDAI.

6.1 Framework for Explainability in XGDAI

6.1.1 Cross-platform explainability

In the current state-of-the-art, explanation techniques, particularly for symbolic deliberative XGDAI, are domain-specific, relying heavily on the domain knowledge (or learning of the domain knowledge) for constructing explanations. In many such applications, agents already have a set of plans and a clear decision system to achieve a goal; thus,

constructing a very rich explanation is often relatively straightforward. However, applying these techniques can be problematic as agents would perform optimally in a specific domain and suboptimal in other domains where knowledge of the agent's world is not fully represented in the framework. A cross-platform explainability framework can significantly benefit existing work in this regard to improve agents' performance and minimize reengineering of the domain knowledge.

Some emerging techniques, such as the domain-agnostics approaches, suggest a useful notion of explainability that enables cross-platform explanation generation for agents and robots. The majority of these techniques can be found for reactive black box agents whose decisions are based solely on the current environment state, not a priori defined. Without a world or domain knowledge model, these agents can generate explanations based on the policy learned. However, these platforms are less extensive, and a significant research effort is still required. A cross-platform explainability approach should significantly benefit both deliberative and reactive agents. Also beneficial would be approaches for bridging the gap between symbolic and reflexive (neural network) approaches to explanation generation and communication.

6.1.2 Theory of Mind for Agent's Teammates

Theory of Mind (ToM) tends to reason about other peoples' perceptions, beliefs, and goals and take them into account [202]. A significant body of work on XGDAI involves agents/robots collaborating with humans and other agents. Given this reality, it is imperative for agents to adequately understand their teammates for effective collaboration and provide a useful and timely explanation when necessary. A useful step in this direction may be to integrate the ToM concept in the explanation framework enabling an agent to also reason about other teammates' perceptions and mental states. A well-constructed ToM should enable the robot to understand its teammates' expectations and thereby provide a useful, relevant, and timely explanation when necessary. The motivation here is that humans are well known to collaborate with their teammates and explain their behavior using the ToM concept extensively. Currently, only one approach mentioned generating explanations about user/team-member expectations and possible violations of expectations [67].

6.1.3 Rich Perceptual-Cognitive Explainability framework

A significant number of approaches (particularly approaches on deliberative XGDAIs) provide an explanation at the level of agents' cognitive functions, i.e., decisions, plans, beliefs, desires, intentions, etc., which are not grounded on actual agents' perception of the real world. On the other hand, a few studies, mainly reactive XGDAIs, highlight procedures for explanation generation at the level of the agent's perceptual function (sensor information, environment states, etc.) with a poor or non-existent explainable cognitive framework or explainable decision-making framework. A rich perceptual-cognitive explainable framework that abstracts low-level agent's perception (primarily perceptual explanatory knowledge) for high-level cognition and explainability would significantly advance the current research on XGDAI. Some emerging techniques of explainability/interpretability would be useful, e.g. Class Activation Map (CAM) to identify the salient information in an image, visual grounding to localize an image component, and referring expression generation (REG) to enable an agent to clearly distinguish between objects in a shared workspace (e.g. in situated human-robot dialogues).

6.1.4 Natural Language Processing

Natural language processing (NLP) is the computer system or AI system's ability to understand, analyze, manipulate, and potentially generate human language [203]. For XGDAI, NLP is crucial in the explanation generation/communication framework to enable a human-comprehensible explanation of agent's/robot's perception, cognition, or decisions. Currently, the state of the art in XGDAI reveals less extensive or even a non-existent natural language processing ability for many of the agents and robots surveyed. The addition of a rich NLP system to existing frameworks would significantly benefit XGDAI in its usefulness and applicability.

6.1.5 Integrating verbal and non-verbal explanation communication

The current state of the art reveals different explanation communication modalities for XGDAI applied in separate niche areas/scenarios. With diverse application scenarios, the need to communicate the agent's/robot's plan, decision, intentions, etc., by combining both verbal and nonverbal communication means is necessary if such an explanation should be natural and effective. A relevant example is seen from how humans explain/communicate using verbal (i.e., speech) and nonverbal (e.g., gesture, facial expression, emotion, etc.), depending on which is most effective for the circumstance. XGDAI should also enable such a combination of different modalities. For example, for agents sharing a pedestrian walkway with humans, rotational head/eye movements seem more natural and effective to communicate the agent's decision to turn left, change paths, etc. [78]. Simultaneously, verbal communication would also prove useful to alert other pedestrians of the agent's approach when they are not aware of its presence. For collaboration

with teammates, speech for normal explanatory conversation is necessary. However, expressive motion like gesture (e.g., head nodding, etc.) would be useful to provide an implicit explanation, or an expressive light [53, 73] or sound to alert teammates of its mental state in an emergency. There can be many possible effective and natural combination of explanation communication modalities. More research work is still required to bridge this gap.

6.1.6   Continual learning of explanatory knowledge

As agents/robots interact with their environment, teammates, and supervisors, they are expected to explain their decisions or actions in different situations and scenarios. Handcrafting explanatory knowledge to satisfy all possible situations and expectations is difficult and requires significant effort or domain engineering. In this respect, an agent's ability to continuously learn to generate/construct explanations in different situations is crucial to explainable agents' success. As in traditional machine learning, learning could be achieved by supervised learning (e.g., learning from demonstration [200]), unsupervised learning [201], and reinforcement learning. Currently, CBR [198], and XRL [33] techniques have been applied in a few studies on deliberative and reactive agents, respectively, to address this concern. However, for reactive XRL, a major concern is how to retain policies learned by agents for constructing explanations. There is still a significant gap to fill in this direction. Issues of scalability and resource management would also need to be addressed if explanatory knowledge is stored in a continual learning framework.

7   Conclusions and further work

The field of XGDAI is emerging with many rich applications in several domains. Explainability enables transparency for these agents/robots and encourages users' trust for applications in safety-critical situations. This survey presents several techniques for explanation generation and communication proposed and implemented in XGDAI until today. Typically, many XGDAIs techniques can enable the robots/agents to justify and explain their decisions and plan rationales for their actions. However, the state of the art shows that current approaches on XGDAIs are still in their infancy, lacking an extensive explanation generation and communication framework. Consequently, this study highlights a framework towards a more extensive XGDAI that has an extended perceptual and cognitive explanation capability. Future work will involve developing a transparent and domain-agnostic integrated architecture for the effective actualization of explainability in XGDAI.

Acknowledgment

Supplementary Materials

Table 1: Some taxonomy on XGDAI. Table presents highlights of key behavioral attributes and application domains of XGDAIs.

Publication	XGDAI Behavior	Application Domain	Trans-parency	Domain specific
[42], [43], [44], [40], [45], [46], [47], [48], [49], [27], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [61], [62], [28], [63], [64], [65], [66]	Deliberative	Robot-human collaborative tasks	Yes	Yes
[67] [68], [16],[69], [70], [71]	Deliberative	Robot-human collaborative tasks	-	No
[72] [205]	Reactive	Robot-human collaborative tasks	Yes	No
[73], [20], [74], [75], [76], [77], [78], [79], [80], [81]	Reactive	Robot-human collaborative tasks	No	No
				Continued on next page

Table 1 – continued from previous page

Publication	XGDAI Behavior	Application Domain	Trans- parency	Domain specific
[82], [40]	Hybrid	Robot-human collaborative tasks	-	-
[66]	Deliberative	Robot navigation	Yes	Yes
[206],	Deliberative	Robot navigation	No	Yes
[208], [83], [20], [84]	Reactive	Robot navigation	No	No
[40], [85], [23]	Hybrid	Robot navigation	-	No
[86]	Deliberative	Game application	-	Yes
[87], [88], [89], [90]	Deliberative	Game application	-	No
[91], [8], [92], [81]	Reactive	Game application	No	No
[93]	Hybrid	Game application	-	No
[66]	Deliberative	Search and Rescue	-	Yes
[94], [95], [96], [97], [98], [99]	Deliberative	Training	Yes	Yes
[100], [101], [52], [51]	Deliberative	E-health	-	-
[102], [103], [104], [105]	Reactive	Ubiquitous computing	No	Yes
[207]	Hybrid	Recommender systems	No	No
[30], [106], [107], [108], [109]	Reactive	Recommender systems	Yes	Yes
[209]	Hybrid	Pervasive systems	No	No
[104], [107]	Reactive	Pervasive systems	No	Yes
[78]	Reactive	Teleoperation	No	No
[45]	Deliberative	Teleoperation	Yes	Yes
[110], [111]	Deliberative	MAS	No	No
[94], [96]	Deliberative	Scenario-based training	Yes	Yes
[210]	Reactive	Monitoring and diagnositcs	No	Yes

References

[1] O. Biran and C. Cotton, "Explanation and justification in machine learning: A survey," in *IJCAI-17 workshop on explainable AI (XAI)*, vol. 8, no. 1, 2017.

[2] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," *Behavioral and brain sciences*, vol. 40, 2017.

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

[4] S. Wermter, S. Griffiths, and S. Heinrich, "Crossmodal lifelong learning in hybrid neural embodied architectures," *Behavioral and Brain Sciences*, vol. 40, p. 72, 2017.

Table 2: Summary of explainability techniques for deliberative XGDAI

Explainability Techniques	References	What is explained?	Key triggers or basis for explanation generation	Transparent/ Post-hoc	Domain-specific/ Domain-agnostic
Goal-driven Autonomy (GDA)	[88], [133], [76]	Plans, goals	Expectation failures, states mismatched	T	DS
Explainable BDI model	[94], [121]	Actions	Underlying belief, desire, and intention	T	DS
SAT	[123]	Agent's current state, goals, plans, reasoning process, and future projections	Underlying beliefs and goals	T	DS
Meta-AQUA Explanation model	[124]	Unusual events or states of the world, agent's reasoning	Anomaly (states mismatched) or interesting event	T	DS
Proactive Explanation	[67]	Agent's decisions	Anticipated surprise or potential expectation violation.		
KAGR	[110]	Agent's reasoning	Events	P	DA
eXplainable-PLAN	[48], [128], [40]	Agent's plan	Plan mismatched	P/T	DA
Explainable NPC	[87], [131]	Agent's actions	On supervisor's request or after active goal completion	P	DA
Debrief	[97]	Agent's decisions	On request or after active goal completion	T	DS
Explainable CREATIVE Model	[50], [132]	Agent's actions (tool creation process)	On request	T	DS

[5] J. Choo and S. Liu, "Visual analytics for explainable deep learning," *IEEE computer graphics and applications*, vol. 38, no. 4, pp. 84–92, 2018.

[6] K. Baum, H. Hermanns, and T. Speith, "From machine ethics to machine explainability and back." in *ISAIM*, 2018.

[7] W. R. Swartout and J. D. Moore, "Explanation in second generation expert systems," in *Second generation expert systems*. Springer, 1993, pp. 543–585.

[8] U. Ehsan, P. Tambwekar, L. Chan, B. Harrison, and M. O. Riedl, "Automated rationale generation: a technique for explainable ai and its effects on human perceptions," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 2019, pp. 263–274.

[9] A. Preece, D. Harborne, D. Braines, R. Tomsett, and S. Chakraborty, "Stakeholders in explainable ai," *arXiv preprint arXiv:1810.00184*, 2018.

[10] F. Puppe, *Systematic introduction to expert systems: Knowledge representations and problem-solving methods*. Springer Science & Business Media, 2012.

[11] A. Bundy, "Preparing for the future of artificial intelligence," *AI & SOCIETY*, vol. 32, no. 2, pp. 285–287, 2017.

Table 3: Summary of explainability techniques for reactive and hybrid XGDAIs

Explainability Techniques	References	What is explained?	Key triggers or basis for explanation generation	Reactive/ Hybrid agents	Transparent/ Post-hoc	Domain-specific/ Domain-agnostic
Automated Rationale Generation (ARG)	[8]	Agent's internal state and action	After each action completes	R	P	DA
Explainable Reinforcement Learning (XRL)	[33], [135], [136], [34], [35]	Agent's decisions, actions or behavior	After active goal completes, perceived user-robot model disparity	R	P	DA
Autonomous Policy Explanation	[74]	Control policies	Inquiry by users	R	P/T	DA
Perceptual-cognitive explanation (PeCoX)	[82], [142], [143] [144]	Agent's behavior	Request of the user	H	P	DA

Table 4: Summary of explanation communication techniques for XGDAIs.

Explanation Communication Types	Techniques	References
Visualization	Class Activation Map (NN Approach)	[145], [146], [147], [148], [149]
	Visual grounding via structured graph (Neural-Symbolic approach)	[150], [151], [152], [153], [154], [155], [156]
	Automata (Neural-Symbolic approach)	[157], [158], [159], [160], [161], [162], [163]
	Rules Extraction Algorithms (Neural-Symbolic approach)	[164]
Numerical Form	Attention Weights	[165], [166], [167], [168], [169], [170]
	SHapley Additive exPlanation	[171]
Textual Communication	Logs	[94], [131]
	Prolog	[50]
Verbal Communication	Speech	[172], [173], [174], [175], [176], [177], [178], [179], [180]
	Referring Expression	[181], [182], [183], [184], [150], [185], [186]
Non-verbal cues	Lights	[53], [73]
	Expressive motion	[78], [187]

[12] B. D. Mittelstadt and L. Floridi, "Transparent, explainable, and accountable ai for robotics," *Science Robotics*, vol. 2, no. 6, 2017.

[13] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling, "Explainable agents and robots: Results from a systematic literature review," in *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 1078–1088.

[14] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[15] A. Holzinger, B. Malle, P. Kieseberg, P. M. Roth, H. Müller, R. Reihs, and K. Zatloukal, "Towards the augmented pathologist: Challenges of explainable-ai in digital pathology," *arXiv preprint arXiv:1712.06657*, 2017.

[16] A. Chandrasekaran, D. Yadav, P. Chattopadhyay, V. Prabhu, and D. Parikh, "It takes two to tango: Towards theory of ai's mind," *arXiv preprint arXiv:1704.00717*, 2017.

[17] J.-C. Park, D.-S. Kim, and Y. Nagai, "Learning for goal-directed actions using rnnpb: Developmental change of "what to imitate"," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 3, pp. 545–556, 2017.

[18] M. Schmerling, G. Schillaci, and V. V. Hafner, "Goal-directed learning of hand-eye coordination in a humanoid robot," in *2015 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. IEEE, 2015, pp. 168–175.

[19] A. D. Dragan, K. C. Lee, and S. S. Srinivasa, "Legibility and predictability of robot motion," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2013, pp. 301–308.

[20] D. Dannenhauer, M. W. Floyd, M. Molineaux, and D. W. Aha, "Learning from exploration: Towards an explainable goal reasoning agent," 2018.

[21] C. M. Kennedy, "A conceptual foundation for autonomous learning in unforeseen situations," in *Proceedings of the 1998 IEEE International Symposium on Intelligent Control (ISIC) held jointly with IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA) Intell*. IEEE, 1998, pp. 483–488.

[22] H. R. Beom, K. C. Koh, and H. S. Cho, "Behavioral control in mobile robot navigation using fuzzy decision making approach," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'94)*, vol. 3. IEEE, 1994, pp. 1938–1945.

[23] Y. Wang, D. Mulvaney, I. Sillitoe, and E. Swere, "Robot navigation by waypoints," *Journal of Intelligent and Robotic Systems*, vol. 52, no. 2, pp. 175–207, 2008.

[24] P. Davidsson, *Autonomous agents and the concept of concepts*. Department of Computer Science, Lund University, 1996.

[25] T. Hellström and S. Bensch, "Understandable robots-what, why, and how," *Paladyn, Journal of Behavioral Robotics*, vol. 9, no. 1, pp. 110–123, 2018.

[26] T. Chakraborti, S. Sreedharan, and S. Kambhampati, "Explicability versus explanations in human-aware planning," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 2018, pp. 2180–2182.

[27] R. H. Wortham, A. Theodorou, and J. J. Bryson, "Improving robot transparency: real-time visualisation of robot ai substantially improves understanding in naive observers," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2017, pp. 1424–1431.

[28] L. Takayama, D. Dooley, and W. Ju, "Expressing thought: improving robot readability with animation principles," in *Proceedings of the 6th international conference on Human-robot interaction*, 2011, pp. 69–76.

[29] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[30] L. Quijano-Sanchez, C. Sauer, J. A. Recio-Garcia, and B. Diaz-Agudo, "Make it personal: a social explanation system applied to group recommendations," *Expert Systems with Applications*, vol. 76, pp. 36–48, 2017.

[31] S. Sreedharan, T. Chakraborti, and S. Kambhampati, "Balancing explicability and explanation in human-aware planning," in *2017 AAAI Fall Symposium*. AI Access Foundation, 2017, pp. 61–68.

[32] K. Knight, "Are many reactive agents better than a few deliberative ones?" in *IJCAI*, vol. 93, 1993, pp. 432–437.

[33] F. Cruz, R. Dazeley, and P. Vamplew, "Memory-based explainable reinforcement learning," in *Australasian Joint Conference on Artificial Intelligence*. Springer, 2019, pp. 66–77.

[34] Z. Juozapaitis, A. Koul, A. Fern, M. Erwig, and F. Doshi-Velez, "Explainable reinforcement learning via reward decomposition," in *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*, 2019, pp. 47–53.

[35] A. Tabrez and B. Hayes, "Improving human-robot interaction through explainable reinforcement learning," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2019, pp. 751–753.

[36] L.-J. Lin, *Self-improving reactive agents: Case studies of reinforcement learning frameworks*. Carnegie-Mellon University. Department of Computer Science, 1990.

[37] L.-J. Lin, "Self-improving reactive agents based on reinforcement learning, planning and teaching," *Machine learning*, vol. 8, no. 3-4, pp. 293–321, 1992.

[38] M. Wooldridge, "Conceptualising and developing agents," in *Proceedings of the UNICOM Seminar on Agent Software*, vol. 42. London, 1995.

[39] A. L. Hayzelden and J. Bigham, *Software agents for future communication systems*. Springer Science & Business Media, 1999.

[40] R. Korpan and S. L. Epstein, "Toward natural explanations for a robot's navigation plans," *Notes from the Explainable Robotic Systems Worshop, Human-Robot Interaction*, 2018.

[41] A. S. Rao, M. P. Georgeff *et al.*, "Bdi agents: from theory to practice." in *ICMAS*, vol. 95, 1995, pp. 312–319.

[42] D. V. Pynadath, N. Wang, E. Rovira, and M. J. Barnes, "Clustering behavior to recognize subjective beliefs in human-agent teams," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 2018, pp. 1495–1503.

[43] M. Sridharan and B. Meadows, "Towards a theory of explanations for human–robot collaboration," *KI-Künstliche Intelligenz*, vol. 33, no. 4, pp. 331–342, 2019.

[44] M. Oudah, T. Rahwan, T. Crandall, and J. W. Crandall, "How ai wins friends and influences people in repeated games with cheap talk," in *Thirty-second AAAI conference on artificial intelligence*. AAAI Press, 2018, pp. 1519–1526.

[45] H. Hastie, F. J. Chiyah Garcia, D. A. Robb, A. Laskov, and P. Patron, "Miriam: A multimodal interface for explaining the reasoning behind actions of remote autonomous systems," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018, pp. 557–558.

[46] Z. Gong and Y. Zhang, "Behavior explanation as intention signaling in human-robot teaming," in *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2018, pp. 1005–1011.

[47] J. Y. Chen, S. G. Lakhmani, K. Stowers, A. R. Selkowitz, J. L. Wright, and M. Barnes, "Situation awareness-based agent transparency and human-autonomy teaming effectiveness," *Theoretical issues in ergonomics science*, vol. 19, no. 3, pp. 259–282, 2018.

[48] R. Borgo, M. Cashmore, and D. Magazzeni, "Towards providing explanations for ai planner decisions," *arXiv preprint arXiv:1810.06338*, 2018.

[49] R. W. Wohleber, K. Stowers, J. Y. Chen, and M. Barnes, "Effects of agent transparency and communication framing on human-agent teaming," in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2017, pp. 3427–3432.

[50] H. Wicaksono and C. S. R. Sheh, "Towards explainable tool creation by a robot," in *IJCAI-17 Workshop on Explainable AI (XAI)*, 2017, p. 63.

[51] F. Kaptein, J. Broekens, K. Hindriks, and M. Neerincx, "The role of emotion in self-explanations by cognitive agents," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 2017, pp. 88–93.

[52] F. Kaptein, J. Broekens, K. Hindriks, and M. Neerincx, "Personalised self-explanation by robots: The role of goals versus beliefs in robot-action explanation for children and adults," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2017, pp. 676–682.

[53] K. Baraka, A. Paiva, and M. Veloso, "Expressive lights for revealing mobile service robot state," in *Robot 2015: Second Iberian Robotics Conference*. Springer, 2016, pp. 107–119.

[54] R. H. Wortham, A. Theodorou, and J. J. Bryson, "What does the robot think? transparency as a fundamental design requirement for intelligent systems," in *Ijcai-2016 ethics for artificial intelligence workshop*, 2016.

[55] S. Rosenthal, S. P. Selvaraj, and M. M. Veloso, "Verbalization: Narration of autonomous robot experience." in *IJCAI*, 2016, pp. 862–868.

[56] A. D. Dragan, S. Bauman, J. Forlizzi, and S. S. Srinivasa, "Effects of robot motion on human-robot collaboration," in *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2015, pp. 51–58.

[57] J. Novikova, L. Watts, and T. Inamura, "Emotionally expressive robot behavior improves human-robot collaboration," in *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2015, pp. 7–12.

[58] S. Li, W. Sun, and T. Miller, "Communication in human-agent teams for tasks with joint action," in *International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems*. Springer, 2015, pp. 224–241.

[59] J. Y. Chen, K. Procci, M. Boyce, J. Wright, A. Garcia, and M. Barnes, "Situation awareness-based agent transparency," Army research lab aberdeen proving ground md human research and engineering, Tech. Rep., 2014.

[60] M. W. Boyce, J. Y. Chen, A. R. Selkowitz, and S. G. Lakhmani, "Effects of agent transparency on operator trust," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, 2015, pp. 179–180.

[61] R. Van den Brule, G. Bijlstra, R. Dotsch, D. H. Wigboldus, and W. Haselager, "Signaling robot trustworthiness: Effects of behavioral cues as warnings," *LNCS*, vol. 8239, pp. 583–584, 2013.

[62] M. Lomas, R. Chevalier, E. V. Cross, R. C. Garrett, J. Hoare, and M. Kopack, "Explaining robot actions," in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 2012, pp. 187–188.

[63] U. Jaidee, H. Muñoz-Avila, and D. W. Aha, "Case-based learning in goal-driven autonomy agents for real-time strategy combat tasks," in *Proceedings of the ICCBR Workshop on Computer Games*, 2011, pp. 43–52.

[64] U. Jaidee, H. Munoz-Avila, D. W. Aha, "Integrated learning for goal-driven autonomy," in *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[65] S. R. Haynes, M. A. Cohen, and F. E. Ritter, "Designs for explaining intelligent agents," *International Journal of Human-Computer Studies*, vol. 67, no. 1, pp. 90–110, 2009.

[66] J. Kröske, K. O'Holleran, and H. Rajaniemi, "Trusted reasoning engine for autonomous systems with an interactive demonstrator," in *4th SEAS DTC Technical Conference. Citeseer*. Citeseer, 2009.

[67] M. T. Gervasio, K. L. Myers, E. Yeh, and B. Adkins, "Explanation to avert surprise." in *IUI Workshops*, vol. 2068, 2018.

[68] T. Chakraborti, S. Sreedharan, Y. Zhang, and S. Kambhampati, "Plan explanations as model reconciliation: Moving beyond explanation as soliloquy," *arXiv preprint arXiv:1701.08317*, 2017.

[69] T. Chakraborti, K. P. Fadnis, K. Talamadupula, M. Dholakia, B. Srivastava, J. O. Kephart, and R. K. Bellamy, "Visualizations for an explainable planning agent," *arXiv preprint arXiv:1709.04517*, 2017.

[70] B. Lettl and A. Schulte, "Self-explanation capability for cognitive agents on-board of ucavs to improve co-operation in a manned-unmanned fighter team," in *AIAA Infotech@ Aerospace (I@ A) Conference*, 2013, p. 4898.

[71] M. Klenk, M. Molineaux, and D. W. Aha, "Goal-driven autonomy for responding to unexpected events in strategy simulations," *Computational Intelligence*, vol. 29, no. 2, pp. 187–206, 2013.

[72] R. T. Chadalavada, H. Andreasson, R. Krug, and A. J. Lilienthal, "That's on my mind! robot to human intention communication through on-board projection on shared floor space," in *2015 European Conference on Mobile Robots (ECMR)*. IEEE, 2015, pp. 1–6.

[73] S. Song and S. Yamada, "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*, 2018, pp. 1–6.

[74] B. Hayes and J. A. Shah, "Improving robot controller transparency through autonomous policy explanation," in *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*. IEEE, 2017, pp. 303–312.

[75] N. Wang, D. V. Pynadath, and S. G. Hill, "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*, 2016, pp. 997–1005.

[76] M. W. Floyd and D. W. Aha, "Incorporating transparency during trust-guided behavior adaptation," in *International Conference on Case-Based Reasoning*. Springer, 2016, pp. 124–138.

[77] I. Shindev, Y. Sun, M. Coovert, J. Pavlova, and T. Lee, "Exploration of intention expression for robots," in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 2012, pp. 247–248.

[78] M. Mikawa, Y. Yoshikawa, and M. Fujisawa, "Expression of intention by rotational head movements for teleoperated mobile robot," in *2018 IEEE 15th International Workshop on Advanced Motion Control (AMC)*. IEEE, 2018, pp. 249–254.

[79] M. Guzdial, J. Reno, J. Chen, G. Smith, and M. Riedl, "Explainable pcgml via game design patterns," *arXiv preprint arXiv:1809.09419*, 2018.

[80] E. Groshev, A. Tamar, M. Goldstein, S. Srivastava, and P. Abbeel, "Learning generalized reactive policies using deep neural networks," *Proceedings International Conference on Automated Planning and Scheduling, ICAPS*.

[81] D. Amir and O. Amir, "Highlights: Summarizing agent behavior to people," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2018, pp. 1168–1176.

[82] M. A. Neerincx, J. van der Waa, F. Kaptein, and J. van Diggelen, "Using perceptual and cognitive explanations for enhanced human-agent team performance," in *International Conference on Engineering Psychology and Cognitive Ergonomics*. Springer, 2018, pp. 204–214.

[83] R. Sukkerd, R. Simmons, and D. Garlan, "Toward explainable multi-objective probabilistic planning," in *2018 IEEE/ACM 4th International Workshop on Software Engineering for Smart Cyber-Physical Systems (SEsCPS)*. IEEE, 2018, pp. 19–25.

[84] A. Jauffret, N. Cuperlier, P. Gaussier, and P. Tarroux, "From self-assessment to frustration, a small step toward autonomy in robotic navigation," *Frontiers in neurorobotics*, vol. 7, p. 16, 2013.

[85] V. G. Santucci, G. Baldassarre, and M. Mirolli, "Grail: a goal-discovering robotic architecture for intrinsically-motivated learning," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 8, no. 3, pp. 214–231, 2016.

[86] H. Munoz-Avila and D. W. Aha, "A case study of goal-driven autonomy in domination games," in *Proceedings of the AAAI Workshop on Goal-Directed Autonomy*, 2010.

[87] M. Molineaux, D. Dannenhauer, and D. W. Aha, "Towards explainable npcs: a relational exploration learning agent," in *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018, pp. 565–569.

[88] M. Molineaux, M. Klenk, and D. Aha, "Goal-driven autonomy in a navy strategy simulation," in *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*. AAAI Press, 2010, p. 1548–1554.

[89] D. Nau, Y. Cao, A. Lotem, and H. Munoz-Avila, "Shop: Simple hierarchical ordered planner," in *Proceedings of the 16th international joint conference on Artificial intelligence-Volume 2*, 1999, pp. 968–973.

[90] D. Voelz, E. André, G. Herzog, and T. Rist, "Rocco: A robocup soccer commentator system," in *Robot Soccer World Cup*. Springer, 1998, pp. 50–60.

[91] J. Andreas, A. Dragan, and D. Klein, "Translating neuralese," *arXiv preprint arXiv:1704.06960*, 2017.

[92] U. Ehsan, B. Harrison, L. Chan, and M. O. Riedl, "Rationalization: A neural machine translation approach to generating natural language explanations," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 81–87.

[93] M. K. Sahota, "Reactive deliberation: An architecture for real-time intelligent control in dynamic environments," in *AAAI*, 1994, pp. 1303–1308.

[94] M. Harbers, K. van den Bosch, and J.-J. Meyer, "Design and evaluation of explainable bdi agents," in *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, vol. 2. IEEE, 2010, pp. 125–132.

[95] J. Broekens, M. Harbers, K. Hindriks, K. Van Den Bosch, C. Jonker, and J.-J. Meyer, "Do you get it? user-evaluated explainable bdi agents," in *German Conference on Multiagent System Technologies*. Springer, 2010, pp. 28–39.

[96] M. Harbers, K. Van Den Bosch, and J.-J. Meyer, "A methodology for developing self-explaining agents for virtual training," in *International Workshop on Languages, Methodologies and Development Tools for Multi-Agent Systems*. Springer, 2009, pp. 168–182.

[97] W. L. Johnson, "Agents that learn to explain themselves." in *AAAI*, 1994, pp. 1257–1263.

[98] S. Kambhampati and S. Kedar, "A unified framework for explanation-based generalization of partially ordered and partially instantiated plans," *Artificial Intelligence*, vol. 67, no. 1, pp. 29–70, 1994.

[99] R. J. Mooney and S. Bennett, "A domain independent explanation-based generalizer." in *AAAI*, 1986, pp. 551–555.

[100] N. C. Codella, M. Hind, K. N. Ramamurthy, M. Campbell, A. Dhurandhar, K. R. Varshney, D. Wei, and A. Mojsilovic, "Teaching meaningful explanations," *arXiv preprint arXiv:1805.11648*, 2018.

[101] A. Grea, L. Matignon, and S. Aknine, "How explainable plans can make planning faster," in *Workshop on Explainable Artificial Intelligence*. Stockholm, Sweden: hal.archives-ouvertes.fr, 2018, pp. 58–64.

[102] B. Y. Lim and A. K. Dey, "Design of an intelligible mobile context-aware application," in *Proceedings of the 13th international conference on human computer interaction with mobile devices and services*, 2011, pp. 157–166.

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

[104] S. Stumpf, W.-K. Wong, M. Burnett, and T. Kulesza, "Making intelligent systems understandable and controllable by end users," 2010.

[105] K. Rehman, F. Stajano, and G. Coulouris, "Visually interactive location-aware computing," in *International Conference on Ubiquitous Computing*. Springer, 2005, pp. 177–194.

[106] M. Nilashi, D. Jannach, O. bin Ibrahim, M. D. Esfahani, and H. Ahmadi, "Recommendation quality, transparency, and website quality for trust-building in recommendation agents," *Electronic Commerce Research and Applications*, vol. 19, pp. 70–84, 2016.

[107] D. Holliday, S. Wilson, and S. Stumpf, "The effect of explanations on perceived control and behaviors in intelligent systems," in *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, 2013, pp. 181–186.

[108] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W.-K. Wong, "Too much, too little, or just right? ways explanations impact end users' mental models," in *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. IEEE, 2013, pp. 3–10.

[109] T. Kulesza, S. Stumpf, M. Burnett, W.-K. Wong, Y. Riche, T. Moore, I. Oberst, A. Shinsel, and K. McIntosh, "Explanatory debugging: Supporting end-user debugging of machine-learned programs," in *2010 IEEE Symposium on Visual Languages and Human-Centric Computing*. IEEE, 2010, pp. 41–48.

[110] A. H. Sbai, W. L. Chaari, and K. Ghedira, "Intra-agent explanation using temporal and extended causal maps," *Procedia Computer Science*, vol. 22, pp. 241–249, 2013.

[111] A. Hedhili, W. L. Chaari, and K. Ghédira, "Explanation language syntax for multi-agent systems," in *2013 World Congress on Computer and Information Technology (WCCIT)*. IEEE, 2013, pp. 1–6.

[112] C. Breazeal, C. D. Kidd, A. L. Thomaz, G. Hoffman, and M. Berlin, "Effects of nonverbal communication on efficiency and robustness in human-robot teamwork," in *2005 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2005, pp. 708–713.

[113] T. Kim and P. Hinds, "Who should i blame? effects of autonomy and transparency on attributions in human-robot interaction," in *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2006, pp. 80–85.

[114] B. Auslander, M. Molineaux, D. W. Aha, A. Munro, and Q. Pizzini, "Towards research on goal reasoning with the tao sandbox," Navy Center for Applied Research in Artificial Intelligence Washington Dc, Tech. Rep., 2009.

[115] R. Reiter and J. De Kleer, "An assumption-based truth-maintenance system," *Artificial Intelligence*, pp. 127–162, 1986.

[116] D. S. Nau, "Current trends in automated planning," *AI magazine*, vol. 28, no. 4, pp. 43–43, 2007.

[117] M. Bratman *et al.*, *Intention, plans, and practical reason*. Harvard University Press Cambridge, MA, 1987, vol. 10.

[118] M. Georgeff, B. Pell, M. Pollack, M. Tambe, and M. Wooldridge, "The belief-desire-intention model of agency," in *International workshop on agent theories, architectures, and languages*. Springer, 1998, pp. 1–10.

[119] F. C. Keil, "Explanation and understanding," *Annu. Rev. Psychol.*, vol. 57, pp. 227–254, 2006.

[120] B. F. Malle, "How people explain behavior: A new theoretical framework," *Personality and social psychology review*, vol. 3, no. 1, pp. 23–48, 1999.

[121] R. Flin and K. Arbuthnot, *Incident command: Tales from the hot seat*. Routledge, 2017.

[122] B. Keysar, S. Lin, and D. J. Barr, "Limits on theory of mind use in adults," *Cognition*, vol. 89, no. 1, pp. 25–41, 2003.

[123] M. R. Endsley, *Innovative model for situation awareness in dynamic defense systems*. CRC Press, 2018.

[124] M. T. Cox, "Perpetual self-aware cognitive agents," *AI magazine*, vol. 28, no. 1, pp. 32–32, 2007.

[125] M. T. Cox and A. Ram, "Introspective multistrategy learning: On the construction of learning strategies," Wright State Univ Dayton OH Dept of Computer Science and Engineering, Tech. Rep., 1999.

[126] S. Sohrabi, J. A. Baier, and S. A. McIlraith, "Preferred explanations: Theory and generation via planning," in *Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

[127] B. Seegebarth, F. Müller, B. Schattenberg, and S. Biundo, "Making hybrid plans more clear to human users—a formal approach for generating sound explanations," in *Proceedings of the Twenty-Second International Conference on International Conference on Automated Planning and Scheduling*, 2012, pp. 225–233.

[128] J. Bidot, S. Biundo, T. Heinroth, W. Minker, F. Nothdurft, and B. Schattenberg, "Verbal plan explanations for hybrid planning." in *MKWI*. Citeseer, 2010, pp. 2309–2320.

[129] S. Biundo and B. Schattenberg, "From abstract crisis to concrete relief – a preliminary report on combining state abstraction and htn planning," in *In Proceedings of the European Conference on Planning*. Springer Verlag, 2001, pp. 157–168.

[130] S. L. Epstein, A. Aroor, M. Evanusa, E. I. Sklar, and S. Parsons, "Learning spatial models for navigation," in *International Conference on Spatial Information Theory*. Springer, 2015, pp. 403–425.

[131] M. Van Lent, W. Fisher, and M. Mancuso, "An explainable artificial intelligence system for small-unit tactical behavior," in *Proceedings of the national conference on artificial intelligence*. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004, pp. 900–907.

[132] N. Lavrac and S. Dzeroski, "Inductive logic programming." in *WLP*. Springer, 1994, pp. 146–160.

[133] B. G. Weber, M. Mateas, and A. Jhala, "Learning from demonstration for goal-driven autonomy," in *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*. AAAI Press, 2012, p. 1176–1182.

[134] N. Block, "Two neural correlates of consciousness," *Trends in cognitive sciences*, vol. 9, no. 2, pp. 46–52, 2005.

[135] M. L. Littman, "Memoryless policies: Theoretical limitations and practical results," in *From Animals to Animats 3: Proceedings of the third international conference on simulation of adaptive behavior*, vol. 3. Cambridge, MA, 1994, p. 238.

[136] P. Sequeira and M. Gervasio, "Interestingness elements for explainable reinforcement learning: Understanding agents' capabilities and limitations," *arXiv preprint arXiv:1912.09007*, 2019.

[137] P. Madumal, T. Miller, L. Sonenberg, and F. Vetere, "Explainable reinforcement learning through a causal lens," *arXiv preprint arXiv:1905.10958*, 2019.

[138] R. Pocius, L. Neal, and A. Fern, "Strategic tasks for explainable reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 10 007–10 008.

[139] R. E. Fikes and N. J. Nilsson, "Strips: A new approach to the application of theorem proving to problem solving," *Artificial intelligence*, vol. 2, no. 3-4, pp. 189–208, 1971.

[140] J. van der Waa, J. van Diggelen, M. A. Neerincx, and S. Raaijmakers, "Icm: An intuitive model independent and accurate certainty measure for machine learning." in *ICAART (2)*, 2018, pp. 314–321.

[141] M. J. Robeer, "Contrastive explanation for machine learning," Master's thesis, 2018.

[142] D. C. Dennett, "Three kinds of intentional psychology," *Perspectives in the philosophy of language: A concise anthology*, pp. 163–186, 1978.

[143] M. Harbers, J. Broekens, K. Van Den Bosch, and J.-J. Meyer, "Guidelines for developing explainable cognitive models," in *Proceedings of ICCM*. Citeseer, 2010, pp. 85–90.

[144] S. A. Döring, "Explaining action by emotion," *The Philosophical Quarterly*, vol. 53, no. 211, pp. 214–230, 2003.

[145] H. Siqueira, S. Magg, and S. Wermter, "Efficient facial feature learning with wide ensemble-based convolutional neural networks," *arXiv preprint arXiv:2001.06338*, 2020.

[146] H. G. Ng, M. Kerzel, J. Mehnert, A. May, and S. Wermter, "Classification of mri migraine medical data using 3d convolutional neural network," in *International Conference on Artificial Neural Networks*. Springer, 2018, pp. 300–309.

[147] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[148] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[149] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *2010 ieee computer society conference on computer vision and pattern recognition-workshops*. IEEE, 2010, pp. 94–101.

[150] M. Shridhar, D. Mittal, and D. Hsu, "Ingress: Interactive visual grounding of referring expressions," *The International Journal of Robotics Research*, vol. 39, no. 2-3, pp. 217–232, 2020.

[151] E. Conser, K. Hahn, C. M. Watson, and M. Mitchell, "Revisiting visual grounding," *arXiv preprint arXiv:1904.02225*, 2019.

[152] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei, "Image retrieval using scene graphs," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3668–3678.

[153] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5410–5419.

[154] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2016, pp. 91–99.

[155] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph r-cnn for scene graph generation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 670–685.

[156] L. Zhou, N. Louis, and J. J. Corso, "Weakly-supervised video object grounding from text by loss weighting and object interaction," *arXiv preprint arXiv:1805.02834*, 2018.

[157] C. L. Giles, C. B. Miller, D. Chen, H.-H. Chen, G.-Z. Sun, and Y.-C. Lee, "Learning and extracting finite state automata with second-order recurrent neural networks," *Neural Computation*, vol. 4, no. 3, pp. 393–405, 1992.

[158] Z. Zeng, R. M. Goodman, and P. Smyth, "Learning finite state machines with self-clustering recurrent networks," *Neural Computation*, vol. 5, no. 6, pp. 976–990, 1993.

[159] C. W. Omlin and C. L. Giles, "Extraction of rules from discrete-time recurrent neural networks," *Neural networks*, vol. 9, no. 1, pp. 41–52, 1996.

[160] G. Weiss, Y. Goldberg, and E. Yahav, "Extracting automata from recurrent neural networks using queries and counterexamples," in *International Conference on Machine Learning*, 2018, pp. 5247–5256.

[161] S. Wermter, "Knowledge extraction from transducer neural networks," *Applied Intelligence*, vol. 12, no. 1-2, pp. 27–42, 2000.

[162] S. Wermter, "Preference moore machines for neural fuzzy integration," in *IJCAI*, 1999, pp. 840–845.

[163] G. Arevian, S. Wermter, and C. Panchev, "Symbolic state transducers and recurrent neural preference machines for text mining," *International journal of approximate reasoning*, vol. 32, no. 2-3, pp. 237–258, 2003.

[164] R. Andrews, J. Diederich, and A. B. Tickle, "Survey and critique of techniques for extracting rules from trained artificial neural networks," *Knowledge-based systems*, vol. 8, no. 6, pp. 373–389, 1995.

[165] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, "Learning to diagnose with lstm recurrent neural networks," *arXiv preprint arXiv:1511.03677*, 2015.

[166] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart, "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism," in *Advances in Neural Information Processing Systems*, 2016, pp. 3504–3512.

[167] Y. Sha and M. D. Wang, "Interpretable predictions of clinical outcomes with an attention-based recurrent neural network," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2017, pp. 233–240.

[168] T. Bai, S. Zhang, B. L. Egleston, and S. Vucetic, "Interpretable representation learning for healthcare via capturing disease progression through time," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 43–51.

[169] J. Lee, J.-H. Shin, and J.-S. Kim, "Interactive visualization and manipulation of attention-based neural machine translation," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2017, pp. 121–126.

[170] S. Liu, T. Li, Z. Li, V. Srikumar, V. Pascucci, and P.-T. Bremer, "Visual interrogation of attention-based models for natural language inference and machine comprehension," Lawrence Livermore National Lab.(LLNL), Livermore, CA (United States), Tech. Rep., 2018.

[171] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, 2017, pp. 4765–4774.

[172] C. Breazeal, A. Takanishi, and T. Kobayashi, *Social Robots that Interact with People.* Springer Berlin Heidelberg, 2008, pp. 1349–1369.

[173] K. Shirai, H. Fujisawa, and Y. Sakai, "Ear and voice of the wabot," *Bull. Sci. & Eng. Research Lab. Waseda Univ*, no. 62, 1973.

[174] K. Shirai and H. Fujisawa, "An algorithm for spoken sentence recognition and its application to the speech input-output system." *IEEE Transactions on Systems, Man, & Cybernetics*, vol. 4, no. 5, p. 475–479, 1974.

[175] T. Kobayashi, "Speech conversation system of the musician robot," *Proc. ICAR'85*, pp. 483–488, 1985.

[176] J. Osada, S. Ohnaka, and M. Sato, "The scenario and design process of childcare robot, papero," in *Proceedings of the 2006 ACM SIGCHI international conference on Advances in computer entertainment technology*, 2006, pp. 80– es.

[177] M. Sato, A. Sugiyama, and S. Ohnaka, "Auditory system in a personal robot, papero," in *2006 Digest of Technical Papers International Conference on Consumer Electronics*. IEEE, 2006, pp. 19–20.

[178] Y. Sakagami, R. Watanabe, C. Aoyama, S. Matsunaga, N. Higaki, and K. Fujimura, "The intelligent asimo: System overview and integration," in *IEEE/RSJ international conference on intelligent robots and systems*, vol. 3. IEEE, 2002, pp. 2478–2483.

[179] R. Nisimura, T. Uchida, A. Lee, H. Saruwatari, K. Shikano, and Y. Matsumoto, "Aska: receptionist robot with speech dialogue system," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 2. IEEE, 2002, pp. 1314–1319.

[180] A. M. Sabelli, T. Kanda, and N. Hagita, "A conversational robot in an elderly care center: an ethnographic study," in *2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2011, pp. 37–44.

[181] L. Yu, H. Tan, M. Bansal, and T. L. Berg, "A joint speaker-listener-reinforcer model for referring expressions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7282–7290.

[182] R. Fang, C. Liu, L. She, and J. Chai, "Towards situated dialogue: Revisiting referring expression generation," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 392–402.

[183] M. Giuliani, M. E. Foster, A. Isard, C. Matheson, J. Oberlander, and A. Knoll, "Situated reference in a hybrid human-robot interaction system," in *Proceedings of the 6th international natural language generation conference*. Association for Computational Linguistics, 2010, pp. 67–75.

[184] R. Dale and E. Reiter, "Computational interpretations of the gricean maxims in the generation of referring expressions," *Cognitive science*, vol. 19, no. 2, pp. 233–263, 1995.

[185] A. Kranstedt and I. Wachsmuth, "Incremental generation of multimodal deixis referring to objects," in *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*, 2005.

[186] M. E. Foster, E. G. Bard, M. Guhe, R. L. Hill, J. Oberlander, and A. Knoll, "The roles of haptic-ostensive referring expressions in cooperative, task-based human-robot dialogue," in *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2008, pp. 295–302.

[187] H. Admoni and B. Scassellati, "Social eye gaze in human-robot interaction: a review," *Journal of Human-Robot Interaction*, vol. 6, no. 1, pp. 25–63, 2017.

[188] J. Townsend, T. Chaton, and J. M. Monteiro, "Extracting relational explanations from deep neural networks: A survey from a neural-symbolic perspective," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 9, pp. 3456–3470, 2019.

[189] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (xai): A survey," *arXiv preprint arXiv:2006.11371*, 2020.

[190] S. O. Arik and Y.-h. Liu, "Explaining deep neural networks using unsupervised clustering," *arXiv preprint arXiv:2007.07477*, 2020.

[191] A. Kori, P. Natekar, G. Krishnamurthi, and B. Srinivasan, "Abstracting deep neural networks into concept graphs for concept level interpretability," *arXiv preprint arXiv:2008.06457*, 2020.

[192] A. d. Garcez, M. Gori, L. C. Lamb, L. Serafini, M. Spranger, and S. N. Tran, "Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning," *arXiv preprint arXiv:1905.06088*, 2019.

[193] C. Esteban, O. Staeck, S. Baier, Y. Yang, and V. Tresp, "Predicting clinical events by combining static and dynamic information using recurrent neural networks," in *2016 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 2016, pp. 93–101.

[194] M. Bramer, *Logic programming with Prolog*. Springer, 2005, vol. 9.

[195] I. Leite, C. Martinho, and A. Paiva, "Social robots for long-term interaction: a survey," *International Journal of Social Robotics*, vol. 5, no. 2, pp. 291–308, 2013.

[196] K. v. Deemter, A. Gatt, I. v. d. Sluis, and R. Power, "Generation of referring expressions: Assessing the incremental algorithm," *Cognitive science*, vol. 36, no. 5, pp. 799–836, 2012.

[197] E. Krahmer and K. Van Deemter, "Computational generation of referring expressions: A survey," *Computational Linguistics*, vol. 38, no. 1, pp. 173–218, 2012.

[198] C. Urdiales, E. J. Perez, J. Vázquez-Salceda, M. Sànchez-Marrè, and F. Sandoval, "A purely reactive navigation scheme for dynamic environments using case-based reasoning," *Autonomous Robots*, vol. 21, no. 1, pp. 65–78, 2006.

[199] J. Kolodner, *Case-based reasoning.* Morgan Kaufmann Publishers Inc., 2014.

[200] M. van Lent and J. Laird, "Learning by observation in a complex domain," in *Proceedings of the Knowledge Acquisition Workshop*, 1998.

[201] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[202] L. J. Byom and B. Mutlu, "Theory of mind: Mechanisms, methods, and new directions," *Frontiers in human neuroscience*, vol. 7, p. 413, 2013.

[203] N. Indurkhya and F. J. Damerau, *Handbook of natural language processing.* Chapman and Hall/CRC, 2010.

[204] P. Langley, B. Meadows, M. Sridharan, and D. Choi, "Explainable agency for intelligent autonomous systems," in *Twenty-Ninth IAAI Conference*, p. 4762, 2017.

[205] X. Gao, R. Gong, Y. Zhao, S. Wang, T. Shu, and S. C. Zhu, "Joint Mind Modeling for Explanation Generation in Complex Human-Robot Collaborative Tasks," in *29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 1119–1126, 2020

[206] X. Pan, T. Zhang, B. Ichter, A. Faust, J. Tan, and S. Ha, "Zero-shot imitation learning from demonstrations for legged robot visual navigation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 679–685, 2020.

[207] J. Xu, H. Wang, Z. Niu, H. Wu, and W. Che, "Knowledge graph grounded goal planning for open-domain conversation generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 5, pp. 9338–9345, 2020.

[208] R. Contreras, A. Ayala, F. Cruz, "Unmanned aerial vehicle control through domain-based automatic speech recognition," *Computers*, vol. 9, no. 3, p. 75, 2020.

[209] F. Alkhabbas, R. Spalazzese, P. Davidsson, "An Agent-Based Approach to Realize Emergent Configurations in the Internet of Things," *Electronics*, vol. 9, no. 9, p. 1347, 2020.

[210] M. Tipaldi, L. Feruglio, P. Denis, G. D'Angelo, "On applying AI-driven flight data analysis for operational spacecraft model-based diagnostics," *Annual Reviews in Control*, vol. 49, pp. 197–211, 2020.