Explainable Goal-Driven Agents and Robots- A Comprehensive Review and New Framework

Fatai Sado[1], Chu Kiong Loo[1]*, Matthias Kerzel[2] Stefan Wermter[2]

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 problems associated with the current generation of artificial intelligence (AI) systems. AI systems particularly dominated by the connectionist deep learning neural network approach lack capabilities of explaining their decisions and actions to others, despite their great successes. They are fundamentally non-intuitive 'black boxes', which renders their decision 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 works on 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 works 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 and understandability, and continual learning for explainability. Finally, the paper presents requirements for explainability and suggests a framework/roadmap for the possible realization of effective goal-driven explainable agents and robots.

Keywords: Explainable AI, Goal-driven agents, Deep neural network, Explainability, Continual learning, Transparency, Accountability.

[1] Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia. e-mail: abdfsado1@gmail.com ckloo.um@um.edu.my

[2] Department of Informatics, Knowledge Technology, University of Hamburg, Vogt-Koelln-Strasse 30, 22527 Hamburg, Germany. e-mail: {matthias.kerzel, wermter}@informatik.uni-hamburg.de

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 [6]. 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.). They engage in tasks that require activity over time, generate plans or goals, and execute their plans in the environment, applying both perception and cognition. They can also adapt the plans or goals as the need arises and may be required to account for their actions [5]. GDAs are useful for many reasons, including scenario-based training (e.g., disaster training), transportation, space and mine exploration, agent development and debugging, and gaming [5]. A relevant example in this context involves an autonomous robot that plans and carries out an exploration task, then participates in a debriefing session where it provides a summary report and answers questions from a human supervisor. According to the observation/recommendation of Swartout and Moore [7], these agents must be able to explain the decisions they made during plan generation, stating alternatives considered; report which actions it executed and why; explain how actual events diverged from the plan and how it adapted in response; and must be able to communicate its decisions and reasons in a human-understandable way. In this review, we focus on two aspects of human-like learning for goal-driven agents and robots: Explainability and Continual lifelong learning. Explainability is enabled by Explainable AI. We focus on both situated and non-situated and embodied/non-embodied autonomous goal-driven eXplainable AI. The review categorizes explanation generation techniques for explainable GDAs according to the agent's perception (e.g., sensory skills, vision, etc.) and cognition (e.g., plans, goals, actions, beliefs, desires, and intentions). It provides a clear taxonomy on eXplainability of GDAs. While an agent's perceptual foundation may be connected to the sub-symbolic reasoning part relating the agents' states, vision, or sensors/environmental information to the agent's cognitive base, the cognitive base relates plans, goals, beliefs, or desire to executed actions . Consequently, we provide a roadmap recommendation for effective actualization of explainable autonomous GDA that has an extended perceptual and cognitive explanation capability.

1.2 What is Explainable AI

Explainable AI refers to artificial intelligence and machine learning techniques that can provide human-understandable justification for their behaviour[1]. Explanations help human collaborators working alongside an autonomous or semi-autonomous agent to understand why an agent failed to achieve a goal or why it completes a task in an unexpected way. 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. In the event that the agent failed to complete the task or performs the task in an unexpected way, it is natural for the human collaborator to want to know why. Explanations thus help the human collaborator understand the circumstances that led to the agent's action, which also allows the collaborator to make an informed decision on how to address that behaviour.

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 [2, 3], the current drive for explainability has been motivated by recent governmental efforts from the European Union, United States (USA) [4], and China [5] which have identified artificial intelligence (AI) and robotics as economic priorities. A key recommendation of the General Data Protection Regulation (GDPR) law of the European Union underlines the right to explanations [6, 7], indicating a must requirement to explain the decisions, actions, or predictions of AI systems. This is to ensure transparency, trust, and users' acceptance of AI systems for safety-critical applications. Thus, pressure is mounting to make AI systems, autonomous agents, and robots transparent, explainable, and accountable[5].

1.4 Data-driven vs Goal-driven XAI

Current industry-led interest in artificial intelligence is almost entirely focussed on data-driven AI. In machine learning, explainability in data-driven AI is often related to the concept of interpretability. A particular system is interpretable if its operations can be understood by a human through introspection or explanation [8]. For instance, Choo and Liu [9] defined the interpretability of a deep learning model as identifying features in the input layer which are responsible for the prediction result at the output layer. Thus, Data-driven XAI implies explaining the decision made by a "black-box" machine learning system, given the data used as input [10]. An important aspect of this branch of XAI is the motivation to find out how available data led to a decision, and whether, given the data and specific circumstances, the machine learning mechanism can consistently reproduce the same decision [11].

On the other hand, there has been limited research on Goal-driven XAI, to date, despite its growing application in an increasingly AI-dependent world [12]. Goal-driven XAI is a research domain that aims at building explainable agents and robots capable of explaining their behaviours to a lay user [6]. These explanations would help the user to build a Theory of the Mind (ToM) of the intelligent agent and would lead to better human-agent collaboration. It would also incite the user to understand the capabilities and the limits of the agents, thereby improving the levels of trust and safety, and avoiding failures. Lack of appropriate mental models and knowledge about the agent may lead to failed interactions [6, 13].

1.5 Emerging Trends in XGDAI

Figure 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 in 2014 and before; however, over the last five years, there is an increasing growth in studies on XGDAI. This upsurge in publication can be seen as the effect of the general pressure on explainability of AI systems and initiatives by several national government agencies like the "right to explanation" by the GDPR [6, 7]. This trend may likely increase exponentially in the upcoming years with several researchers working on XAI in different research domains.

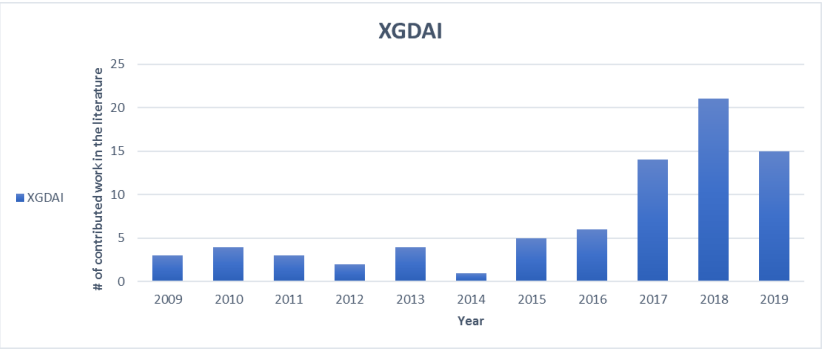

Fig. 1. Emerging trend in XGDAI

2 Terminologies Clarification

2.1 Terms in XGDAI

Different terminologies can be found in the literature for the description of a Goal-driven AI. Some authors used terms such as goal-directed agents or robots [14-17], goal-seeking agents [18-20], or simply autonomous agents[21]. The underlying attribute of these agents is that they seek to achieve a goal or execute a plan. Also, just as in data-driven XAI, several terminologies are used interchangeably for XGDAI, even though there is an acknowledgment for the need of clear taxonomy. Terms such as understandability [22], explicability [23], transparency [24], predictability [16], readability [25], and legibility [16] can be found. In this section, we clarify the distinction and similarities among these terms.

Understandability: denotes the characteristic of an agent to make a human understand its function – how it works – without any need for explaining its internal structure or the algorithm by which the agent/robot processes its data internally[26] [18]

Explicability: According to Sreedharan and Kambhampati [27], an explicable system is one that avoids the need to provide explanations by generating plans that match the human's expected plan.

Explainability: Explainability can be understood as an agent's capability of making decisions that are comprehensible to humans. The explanation acts as an interface between the artificial decision maker and the human [17].

Transparency: transparency refers to the ability to describe, inspect and reproduce the mechanisms through which an AI agent/robot makes decisions and learns to adapt to its environment [26].

Predictability: According to Dragan, et al. [16], predictability refers to the quality of an agent's/robot's behavior or action matching expectations, implying that an agent's/robot's action towards a goal is predictable if it matches what an observer would expect. This, 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 behavior can enable an observer to infer its intention. A legible robot motion can be seen, therefore, as one that enables an observer to quickly and confidently infer the correct goal of the robot [16].

Readability: readability in XGDAI implies the notion of robot behavior to be human-readable such that people can figure out what the robot is doing and can reasonably predict the robot's next action to interact with the robot in an effective way[25].

The overall use of terminologies for XGDAI suggests a trend towards *explicit* explainability where the agent/robot provides a clear explanation for its behavior – decisions or actions – and *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 schemes and react to changes in the agents' environment in a stimulus-response fashion, no model of the world is required (the robot choose one action at a time); and hybrid which combines the above two behaviors [21]. Some other terminologies include goal-driven autonomy, goal-driven agency, and BDI. Table 1 presents taxonomies of XGDAI behavior that can be found in literature. In this section, we make further clarification on these attributes.

Reactive: Reactive agents exhibit a collection of simple behavioral schemes which react to changes in the environment [21], no model of the world is included. They can reach their goal only by reacting reflexively on 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 the use of symbolic representation and manipulation as a base of artificial intelligence [28]. Model-free (deep) reinforcement learning is one of the state-of-the-art approaches that enables reactive agent behavior. Some notable works include MXRL [29], minimal sufficient explanation (MSX) via Reward Decomposition[30], and RARE [31]. The reader is directed to reference [] for more comprehensive literature on reactive RL agents.

Deliberative: Deliberative agents behave more like they are thinking, by searching through a space of behaviors, maintaining an internal state, and predicting the effects of their actions. They plan ahead to reach their goals. Wooldridge defines such agents as "one that possesses an explicitly represented, symbolic model of the world, and in which decisions (for example, about what actions to perform) are made via symbolic reasoning" [32]. According to the traditional approach, the cognitive component of these agents consists of essentially two parts: a planner and a world model [21]. The world model is an internal description of the agent's external environment and sometimes including itself. The planner uses this description to make a plan of how to accomplish the agent's goal. These agents' way of working can be described as a sense-model-plan-act cycle. Most commonly used architecture for implementing such behavior is the Belief-Desire-Intention (BDI) model, where an agent's

beliefs about the world (its image of a world), desires (goal) and intentions are internally represented, and practical reasoning is applied to decide which action to select[33].

Hybrid: There has been considerable research focused on integrating both reactive and deliberative agent strategies resulting in developing a compound called hybrid agent, which combines extensive manipulation ofwith nontrivial symbolic structures and reflexive, reactive responses to external events [21]. This integration of flexibility and robustness of reactivity and the foresight of deliberation is suggested to be the modern drive [34], to integrate the flexibility and robustness of reactivity with the foresight of deliberation. Supporters of the hybrid approach believe it is preferable since both high-level deliberative reasoning and low-level reaction on perceptual stimuli seem necessary for actualizing an agent's behavior. An example is the hybrid system proposed by Wang, et al. [20] that uses reactive exploration to generate waypoints that areis then used by a deliberative system to plan future movements through the same environment. Another existing system with mixed reactive and deliberative behaviors is the agent developed by Rao and Georgeff [35] , which reasons about when to be reactive and when to follow goal-directed plans.

XGDAI Behavior	**References**
Deliberative	Chapman [36], Johnson [37], Kambhampati and Kedar [38], [39], [40], [41],[42], [43],[44],[45],[46],[13],[47],[48],[49], [50],[51],[52],[53],[54],[55],[56],[57],[58],[59], [60],[61], [62], [63], [43], [64], [65], [66],[67], [68]
Reactive	[69], [70],[71],[1],[72],[73],[74],[75],[76],[77],[78], [79], [80], [54], [17], [81], [82], [83], [84], [85], [86], [87],
hybrid	[88], [89], [90], [34], [91], [92], [59]

Table 1. Behavioral attributes of XGDAI in literature

2.3 Application Scenarios for XGDAI

This section presents the application scenarios for XGDAI that are primarily reported in the literature. As presented in Table 2, XGDAI application scenarios include: robot-human collaborative tasks, robot navigation, game applications, search and rescue, training, E-health, ubiquitous computing and recommender systems.

Robot-human collaborative tasks: Tasks such as working closely with humans in a factory setting 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 has been shown to improve the quality of teamwork [93] and to enable both robots or humans to take responsibility (credit or blame) for their actions in collaborative situations [94].

Robot navigation: In robot navigation, Korpan and Epstein [34] proposed a "Why-Plan" that compares the perspectives of an autonomous robot and a person when they plan a path for navigation. A goal-discovering robotic architecture (in a simulated iCub robot) was proposed in [95] to autonomously explore the world and learn different skills that allow the robot to modify the environment.

Game Application: In-game applications, explanations were provided for the non-player characters to reduce the frustration of the human players [96].

Search and Rescue: A search and rescue scenario was implemented for a robot that is tasked with searching the environment and escorting people to the nearest exit [57].

Training: Explainable agents were proposed in a virtual training system for complex, dynamic tasks in which fast decision making is required, like in a search and rescue mission or fire incident [60].

Ubiquitous computing: In ubicomp systems, intelligibility was proposed to allow users to understand how the system works and to let users intervene when the system makes a mistake [97].

E-health: In e-health, explanations that take into account their own and other's emotions were proposed for a PAL (a Personal Assistant for a healthy Lifestyle) agent that interacts with the children, their parents, and their caregivers, to assist them with the treatment process [53].

Recommender systems: Explanation facility for a recommender system called Personalized Social Individual Explanation approach (PSIE) was proposed for group recommendation in [98]. For movie and music recommender systems, explanations were proposed in [99] to investigate how varying soundness and completeness impacted users' mental models.

XGDAI Application scenarios	**References**
Robot-human collaborative tasks	[93], [94], [52], [100], [101], [59], [34], [102], [103], [104], [43], [45], [13], [47], [80], [54], [105], [48], [105], [106], [55], [56],[57], [107], [59], [60], [61], [62], [63], [82], [83], [85], [86], [87], [25], [43], [67],
Robot navigation	[20], [95], [108], [96], [34], [109], [110], [103], [58], [17],
Game Application	[111], [96], [71], [112],
Search and Rescue	[57]
Training	[71],[68],[60],[53]
E-health	[113],[46],[53],
Ubiquitous computing	[114],[97],[115],
Recommender systems	[98],[116], [117], [99], [118],
Pervasive systems	[115], [117],
Teleoperation	[84], [102],
MAS	[49], [50],
Scenario based training	[65],[68]

Table 2. Application scenarios for XGDAI

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 explainable goal-driven AI. The section is further divided into two subsections. The first subsection presents techniques that are applied for deliberative XGDAI, and the second subsection discusses techniques that are applied to reactive XGDAI. The overview of explainability techniques for XGDAI shows that the techniques are either domain- specific, agnostic or post-hoc. Domain- specific explainability techniques are heavily dependent on the domain knowledge of the agent world 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 enable explanations without necessarily tracing the actual reasoning process that led to the decision [6]. A summary of the findings is presented in Table 3.

3.1 Deliberative XGDAI

3.1.1 Models of Goal-Driven Autonomy - *Transparent domain-specific*

Goal-driven autonomy is a conceptual model of goal reasoning which enables an agent to continuously monitor its current plan's execution, assess whether the encountered states match expectations, and generate an explanation on identifying a mismatch [91]. The model enables the agent to determine which goals to pursue, to identify when to select new goals, and to explain why new goals should be pursued [89]. For most of the existing works in goal-driven autonomy, an explanation is generated when an active goal is completed or when a discrepancy is detected between the actual event and the intended plan. Such discrepancies may result if the domain knowledge is flawed, i.e., if the dynamics according to which the state was projected were incorrect or the perception of a state is incorrect. Discrepancies may also result if there is a hidden factor influencing the state, in which case, an explanation is generated to explain the discrepancies and to find or address the hidden factors that affect the state [91].

Molineaux, et al. [91] presented ARTUE (Autonomous Response to Unexpected Events), a GDA domain-independent autonomous agent that dynamically reasons about what goals to pursue in response to unexpected circumstances in a dynamic environment (Fig. 2). ARTUE is implemented in a Sandbox simulation scenario [119]. The ARTUE system integrates four components: a novel Hierarchical Task Network (HTN) planner that reasons about exogenous events by projecting future states in the dynamic continuous environment; an explanation component that reasons about hidden information in the environment by abduction over the conditions and effects found in the planning domain using an Assumption-based Truth Maintenance System [120]; a component that uses domain knowledge in the form of principles to reason about and generate new goals; and a goal management component responsible for prioritizing and issuing goals to the planner. Based on this approach, ARTUE could handle challenges from new and unobservable objects within planning.

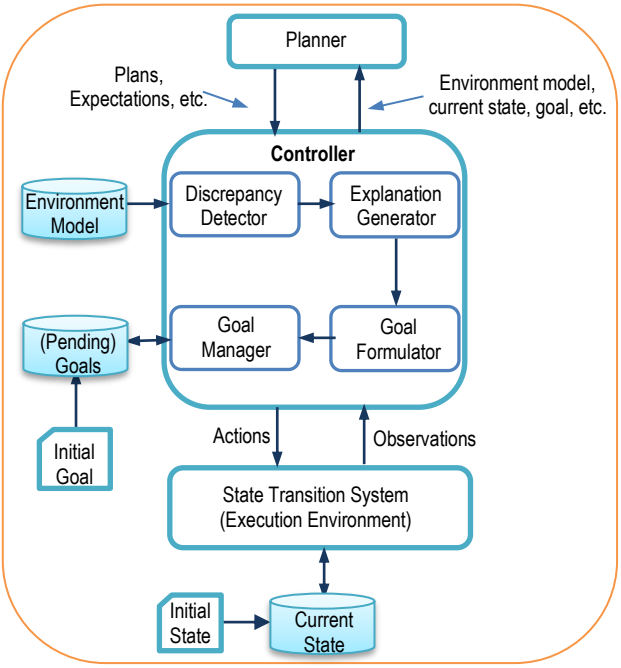

Fig. 2: Conceptual Model of goal-driven autonomy [91]

3.1.2 Explainable BDI model - *Transparent model specific*

Symbolic AI, such as BDI-agents with built-in Beliefs, Desires and Intentions, provide good opportunities for the generation of explanations that are understandable and useful to a human team-member [65]. Harbers, et al. [65] present a model for explainable BDI agents, which enables the explanation of a BDI agent's behaviour or mental state in terms of its underlying beliefs and goals. The motivation is that humans explain and understand their behaviour in terms of underlying desires, goals, beliefs, intentions and the likes [121, 122]. When using BDI agents, the mental concepts underlying an action can be used to explain the action since the agents determine their actions by a deliberation process on their mental concepts. Since mental reasoning is expressed symbolically, the explanation generation process is essentially straightforward. Typically, a behaviour log stores all past mental states and actions of the agent that may be needed for explanations. When there is a request for an explanation, the explanation algorithm is applied to the log, selecting the beliefs and goals that become part of the explanation. However, not all 'explaining elements' can be useful in the explanation [121].

An important aspect of explainable BDI agents is that they can clarify typical human errors. According to Flin and Arbuthnot [123], explainable BDI agents can make trainees aware of their (false) assumptions about other agents' mental states by revealing the agents' actual ones. In many critical situations, people can make false assumptions about the knowledge and intentions of others [123], a phenomenon of attributing incorrect mental states to others [124].

3.1.3 **Situation Awareness–Based Agent Transparency (SAT) Model**

- Transparent model specific

SAT is a model of agent transparency to support operator situation awareness (SA) [125] of the mission environment involving the agent [48, 105]. The SAT model is implemented as a user interface to provide three levels of transparency: the robot's status (e.g., current state, goals, plans), the robot's reasoning process, and the robot's projections (e.g., future environment states). At the first level of the SAT model, an operator is provided with the basic information about the agent's current state and goals, intentions, and proposed actions. At the second level, the operator is provided with information about the agent's reasoning process behind those actions and the constraints/affordances that the agent considers when planning those actions. At the third level, the operator is provided with information regarding the agent's projection of the future state, such as predicted consequences, likelihood of success/failure, and any uncertainty associated with the projections. An agent's transparency in this context is defined as the descriptive quality of an interface pertaining to its abilities to afford an operator's comprehension about an intelligent agent's intent, performance, future plans, and reasoning process. Whereas, an operator's trust is defined as: "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [4].

3.1.4 **Meta-AQUA Explanation model-** ***Transparent domain-specific***

Meta-AQUA is an introspective multi-strategy learning system proposed by Cox [126] that improves its story-understanding performance through a metacognitive analysis of its reasoning failures. It is designed to integrate cognition (planning, understanding, and learning) and metacognition (control and monitoring of cognition) with intelligent behaviours. Meta-AQUA is implemented in the *in*itial in*tro*spective cognitive (INTRO) agent and simulated in the Wumpus World simulated environment, an environment that is partially observable (Fig. 3). The INTRO agent determines its own goals by interpreting and explaining unusual events or states of the world. It couldcan perform actions to change the environment, including turn, move ahead, pick up, and shoot an arrow. As the agent navigates the environment, Meta-AQUA natural language performance task is to "understand" stories by building causal explanatory graphs to link individual events that form the stories. An example of a story can be "S1: The Agent left home," "S2: She traveled down the lower path through the forest.," "S3: At the end she swung left. S4: Drawing her arrow she shot the Wumpus," "S5: It screamed," "S6: She shot the Wumpus, because it threatened her." To understand a story completely, Meta-AQUA explains unusual or surprising events and link them into a causal interpretation that provides the motivations and intent supporting the actions of characters in the story. The approach may also save computational resources by excluding explanations for regular events. Meta-AQUA uses case-based knowledge representations implemented as frames tied together by explanation-patterns to represent general causal structures[126, 127].

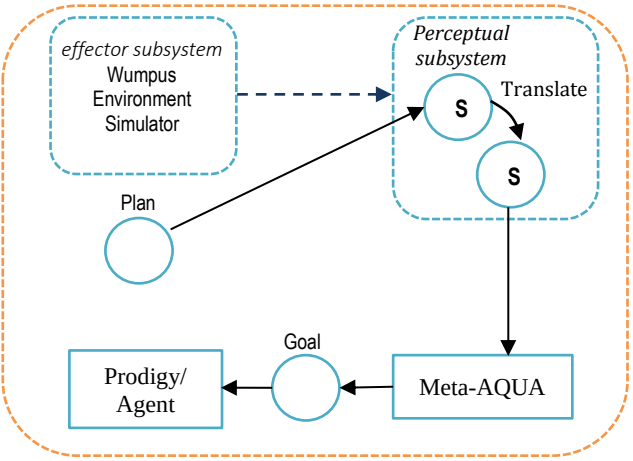

Figure 3. INTRO architecture [126]

3.1.5 **Proactive Explanation model**

In scenarios that involve teams of humans and autonomous agents, a proactive explanation that anticipates or pre-empts potential surprises can be useful. By providing timely explanations, autonomous agents could avoid perceived faulty behaviour and other trust-related issues to enable effective collaboration with team members. Gervasio, et al. [106] presented a framework that uses explanations to avert surprise, given a potential expectation violation. Surprise is used here as the primary motivation for proactivity. When an agent violates expectations, typically, a human collaborator would want to know the reason why it behaved in an unexpected way. In reacting to the human's surprise, the agent tries to explain away the violation by providing an explanation [106].

3.1.6 **Explanation-based generalization (EBG) -** ***Post-hoc domain-specific***

Explanation-based generalization (EBG) is a technique proposed by Mitchell, et al. [128] to formulate generalizations from a single positive training example by constructing explanations. The key idea of EBG is that it is possible to form a justified generalization from a single positive training example provided the learning system (or agent) is endowed with some explanatory capabilities. In particular, the system must be able to explain to itself why the training example is an example of the concept under study. This implies that the system's generalizer should possess a definition of the concept under study as well as the domain knowledge for constructing the required explanation. Generalization involves observing a set of training examples of some general concept, identifying the essential features common to those examples, then formulating a concept definition based on the common features. Earlier techniques, aside from the EBG, apply a generalization method that focuses on generalization from a large number of training examples, employing inductive bias to search for features that are common to the training examples. Unlike EBG, these techniques do not use domain-specific knowledge. They apply a 'black-box' approach that presents some difficulty in justifying the generalizations that they produce. The notion of EBG is to constrain the search by relying on knowledge of the task domain and of the concept under study. EBGs analyse the training example by first constructing an explanation of how the example satisfies the definition of the concept under study and then produce a

valid generalization of the example along with a deductive justification of the generalization in terms of the system's knowledge.

Kambhampati and Kedar [38] extend the EBG method to create and generalize partially ordered and partially instantiated (POPI) plans for agent planning. The approach provides EBG with explanations of the correctness of POPI plans based on Modal Truth Criteria [3], which state the necessary and sufficient conditions for ensuring the truth of a proposition at any point in a plan. The explanations are represented by a set of dependency links, with well-defined semantics, called validations. These explanations are then used as the basis for generalization.

Another fully implemented system designed to generalize the structure of explanations and to produce recursive concepts when warranted is BAGGER2 [41]. This system is the successor to an earlier structure-generalizing explanation-based learning (EBL) system, BAGGER [129]. BAGGER learns iterative concepts (manifested as linear chains of rule applications). BAGGER2 generalizes explanation structures by looking for repeated inter-dependent sub-structures in an explanation. Unlike its predecessor, BAGGER2 can acquire recursive concepts involving arbitrary tree-like applications of rules, it can perform multiple generalizations in one example, and it can integrate the results of multiple examples. BAGGER2 extends the EGGS algorithm proposed in [39], a standard EBL algorithm. Both algorithms assume that, in the course of solving a problem, a collection of pieces of general knowledge (e.g., inference rules, rewrite rules, or plan schemata) are interconnected, using unification to ensure compatibility. Explanation-based learning systems must generalize explanation structures if they are to be able to fully extract general concepts inherent in the solutions to specific examples.

3.1.10 **KAGR Explanation System –** ***Post-hoc domain agnostic***

KAGR is an explanation structure proposed by Sbaï, et al. [49] to explain agent reasoning in a multi-agent systems (MAS), operating in uncontrollable, dynamic or complex situations where an agents' reasoning is not clearly reproducible for the users. KAGR describes the reasoning state of an agent as a tuple <K, A, G, R> (Knowledge, Action, Goal, Relation), in relation to the execution of an agent's action at runtime. Under the KAGR explanation framework, events produced by the agents at runtime are first intercepted, and an explanatory knowledge acquisition phase is performed where knowledge attributes or details related to the execution of detected events are presented in a KAGR structure [49, 50]. A second step is performed where the semantic links between these attributes are expressed in an extended causal map (CM) model that constitutes knowledge representation formalism. In a final step, a natural language interpretation for the CM model is achieved using predicate first-order logic to build up a knowledge-based system for comprehensible explanation to users. Sbaï, et al. [49] adopts a three-module approach to the explainability of a multi-agent system (Fig. 4). The first module generates explanatory knowledge. The second one represents the knowledge in the extended causal maps formalism. The third one then analyses and interprets the built causal maps using a first-order logic to produce reasoning explanations.

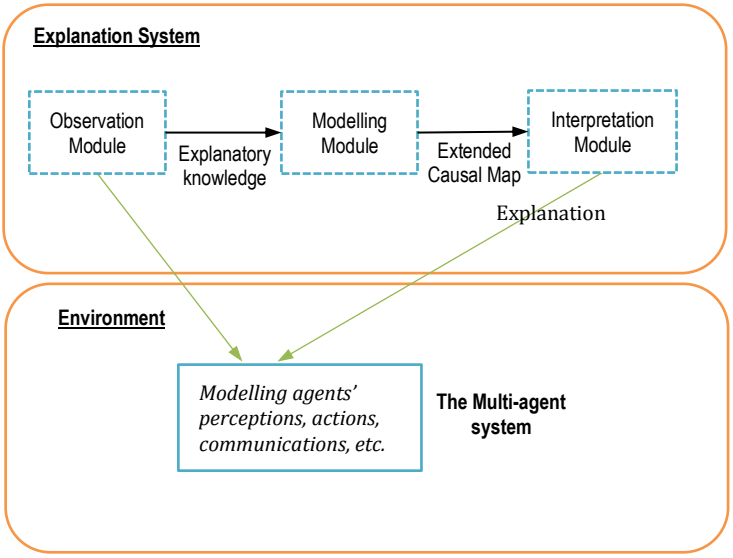

Figure 4. KAGR explanation architecture[49]

3.1.11 eXplainable Plan Models – *Post-hoc/Transparent domain-agnostic*

Plan Explanation is an area of planning where the main goal is to help humans understand the plans produced by the planners (e.g., [130], [131], [132]). This involves the translation of agents' plans to a form that humans can easily understand and the design of interfaces that help this understanding. Relevant works in this context include XAI-PLAN [51], RBP [132], and WHY-PLAN [34].

XAI-PLAN is an explainable plan model proposed by Borgo, et al. [51] to provide initial explanations for the decisions made by an agent planner [51]. Under the framework, explanations are created by allowing the user to explore alternative actions in the plans and then compare the resulting plans with the one found by the planner. The interaction between the user and the planner enhances mixed-initiative planning that have a likelihood to improve the final plan. The XAI-PLAN methodology is domain-independent and agnostic about the planning system. XAI-PLAN provide answers to questions like, "*why does the plan contain action A rather than action B (that I would expect)?"*. The algorithm takes as input an initial set of plans; a user selects an action in the plan; the XAI-Plan node implements the algorithms for generating explanatory plans and communicates to the user through a user interface (Fig. 5). The knowledge base, problem interface, and planner interface are supplied by ROSPlan, which are used to store a Planning Domain Definition Language (PDDL) model and provide an interface to the AI planner, i.e., an architecture for embedding task planning into ROS systems.

Refinement-based planning (RBP) is a transparent domain-independent framework proposed by Bidot, et al. [132] to enable verbal human queries and to produce verbal plan explanations. RBP allows for an explicit representation of the search space explored during the plan generation process; giving the possibility to explore the search space backwards to search for the relevant flaws and plan modifications. RBP is based on a hybrid planning framework that integrates partial-order causal-link planning and hierarchical planning [133],

using states and action primitives. RBP is implemented in PANDA, which integrates hierarchical planning and partial-order causal-link planning. The human user inputs a set of partially ordered tasks and asks for explanations that can justify the ordering of two tasks or the temporal position of a task.

Korpan and Epstein [34] proposed **Why-Plan** as an explanation method that compares the perspectives of an autonomous robot and a person when they plan a path for navigation. 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. The framework suggests that a cognitive basis for a robot controller facilitates the production of natural explanations, i.e., a controller that uses human-like rationales to make decisions can readily produce natural explanations. Why-Plan is implemented in SemaFORR, a cognitively-based hybrid robot controller that learns a human-like spatial model [134], and makes a decision using two sets of reactive Advisors. The first Advisor, the rule-based Advisors to mandate clearly correct actions or veto unacceptable ones, and the second, common-sense Advisors to vote to select an action if no action has been mandated and multiple choices remain. SemaFORR demonstrates **Why-Plan's**ability to produce meaningful, human-friendly explanations quickly in natural language.

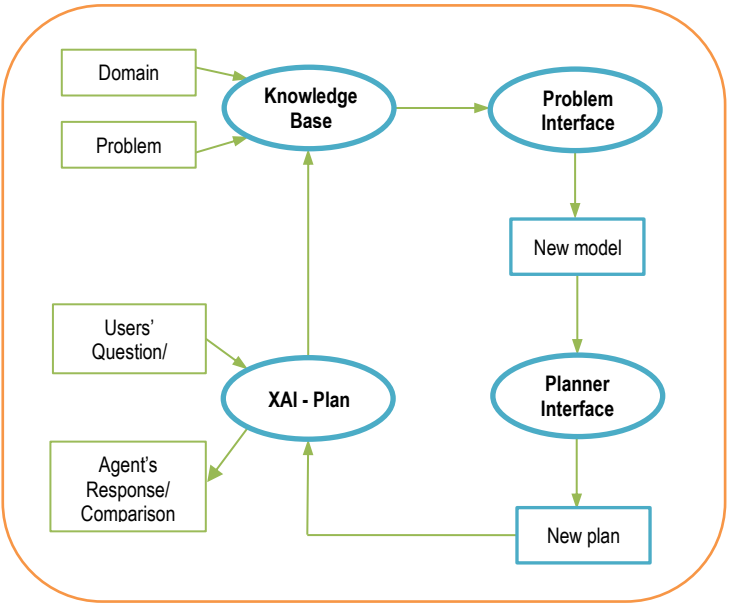

Figure 5: XAI – Plan architecture [51]

3.1.12 **Explainable NPC** – *Post-hoc domain agnostic*

Explainable Non-Player Characters (Explainable NPC) is an architecture proposed by Molineaux, et al. [96] to minimize the frustration of video game players by providing an explanation for NPC actions. To many video game players, non-player characters (NPCs) can be a major source of frustration because of the opacity of their reasoning process. The NPC may be responding to internal needs that a player is unaware of or encounter obstacles

that a player cannot see, but they do not communicate these problems [96]. The Explainable NPC architecture is thus motivated to enable agents to learn about their environments, accomplish a range of goals, and explain what they are doing to a supervisor (Fig. 6). The agent receives an observation at each time step that reflects information about the true environment state and interacts with the environment by taking an action. The agent also interacts with a supervisor. The supervisor makes requests to the agent, updated at each time step, that reflect what the supervisor would like the agent to accomplish. In return, the agent is expected to provide an explanation to the supervisor at each time step describing why it takes a particular action. The framework also describes an evaluation technique centred around the supervisor's satisfaction and understanding of the agent's behaviour. The framework (Fig. 6) is divided into four submodules: the exploratory planner, responsible for taking actions to obtain new information with which to update an action model, the goal-directed planner, responsible for achieving goals given by the supervisor, the transition model learner, responsible for updating the agent's model of the world, and the controller, responsible for determining when to explore, achieve goals, and update the model as well as communicating with the supervisor [96].

In a related work, Van Lent, et al. [135] proposed the eXplainable AI (XAI) architecture for NPCs in a training system. The Explainable AI (XAI) works during the after-action review phase to extract key events and decision points from the playback log and consequently allow the non-player AI-controlled characters to explain their behaviour in response to the questions selected from the XAI menu. The NPC AI is divided into two AI subsystems; a Control AI and a Command AI. The Control AI is responsible for reactive behaviour and low-level actions of individual characters, whereas the Command AI generates higher-level behaviour. The Explainable AI system in Full Spectrum Command logs the activities of the NPC AI system during the execution phase and uses that log during the after-action review phase. The log consists of a long sequence of AI events records, each with an associated timestamp.

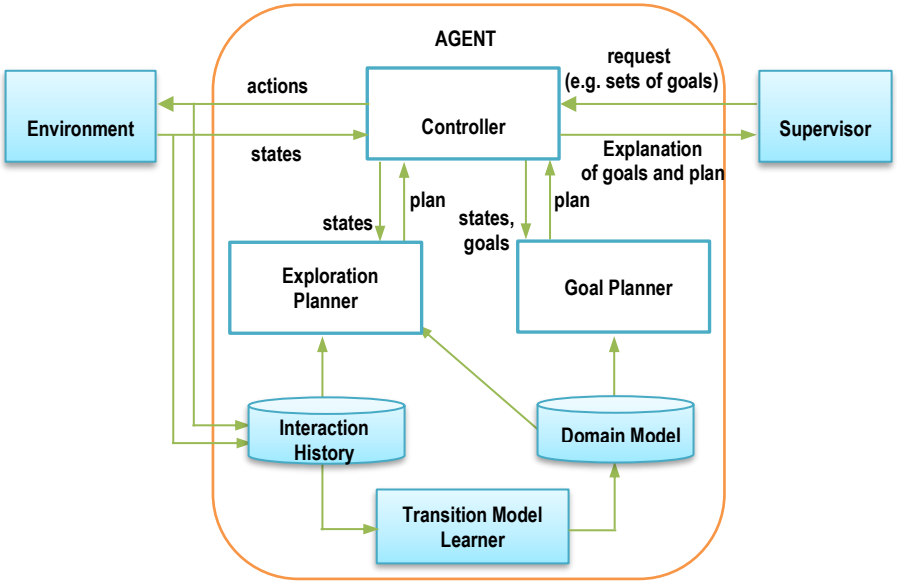

Figure 6: Explainable NPC Architecture [96]

3.1.14 **Debrief** – *Transparent domain-specific*

Debrief is a multimedia explanation system proposed by Johnson [37] that constructs explanations by recalling the situation in which the decision was made and by replaying the decision under variants of the original situation or through experimentation to determine what factors were critical for the decision. The factors are critical in the sense that if they were not present, the outcome of the decision process would have been different. Details of the agent's implementation, such as which individual rules were applied in making the decision, are automatically filtered out. It is not necessary to maintain a complete trace of rule firings to produce explanations. The relationships between situational factors and decisions are learned so that they can be applied to similar decisions.

3.1.15 **Explainable CREATIVE Model** – *Transparent domain-specific*

Cognitive Robot Equipped with Autonomous Tool Invention Expertise (CREATIVE) is a relational approach proposed by Wicaksono and Sheh [43] to enable a robot to learn how to use an object as a tool and, if needed, to design and construct a new tool. To get explanations, or to make the tool creation explainable for a human, all relevant information is stored, including the learned hypotheses, in Prolog. CREATIVE utilises relational representation, so its results are inherently explainable as it describes the relation between objects as Prolog facts. The relational representations of tool models are learned by a form of Inductive Logic Programming (ILP) [136].

The robot starts without a complete action model, so it cannot construct a plan. The robot can learn by observing a single correct example given by a tutor and build an initial novel model to complete the previous ones. A problem solver must achieve a goal and may use a tool to do so. The results of its attempt to solve the problem are sent to a critic that determines if they correspond to the expectations of a planner (Fig. 7). Depending on that assessment, learning by trial and error is performed, via an ILP learner, to update a relevant action model. A problem generator selects a new experiment to test the updated model. If there is no suitable tool to accomplish the task, tool invention is performed. A label from a critic is passed to the tool generalizer and manufacturer, and via a simple user interface, a human user can get explanations from the robot.

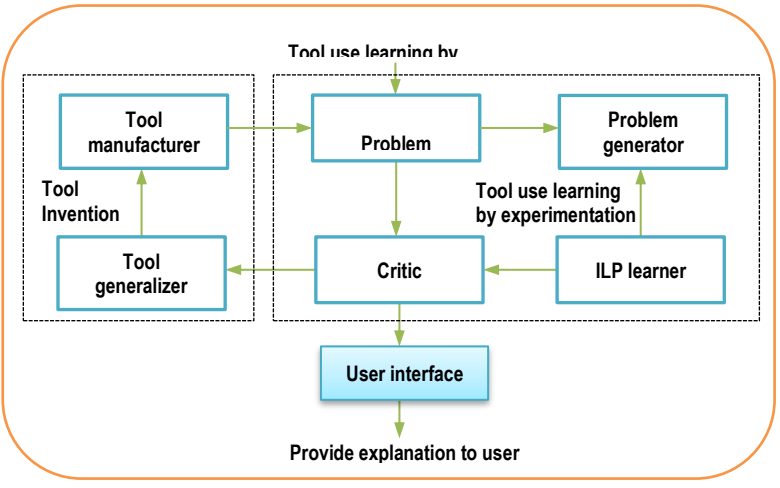

Figure 7: Explainable tool creation to the user

3.2 Reactive XGDAI

3.2.1 **Transparent Reactive Planning Technique** – *Post-hoc domain-agnostic*

Wortham, et al. [80] proposed the instinct reactive planner as a transparent action selection mechanism for a low-cost ARDUINO-based maker robot named robot R5 [80]. The approach is to use reactive planning techniques to build transparent autonomous agents. The Instinct Planner includes capabilities to facilitate plan design and runtime debugging (Fig. 8). It reports the execution and status of every plan element in real-time, allowing to implicitly capture the reasoning process within the robot that gives rise to its behaviour. The planner can report its activity as it runs using call-back functions to a monitor class. Six separate call-backs monitor the Execution, Success, Failure, Error and In-Progress status events, and the Sense activity of each plan element. In the R5 robot, the call-backs write textual data to a TCP/IP stream over a wireless (wifi) link, and a JAVA based Instinct Server receives this information and logs the data to disk. This communication channel also allows for commands to be sent to the robot while it is running. Figure 10 shows the overall architecture of the instinct planner within the R5 robot, communicating via wifi to the logging server.

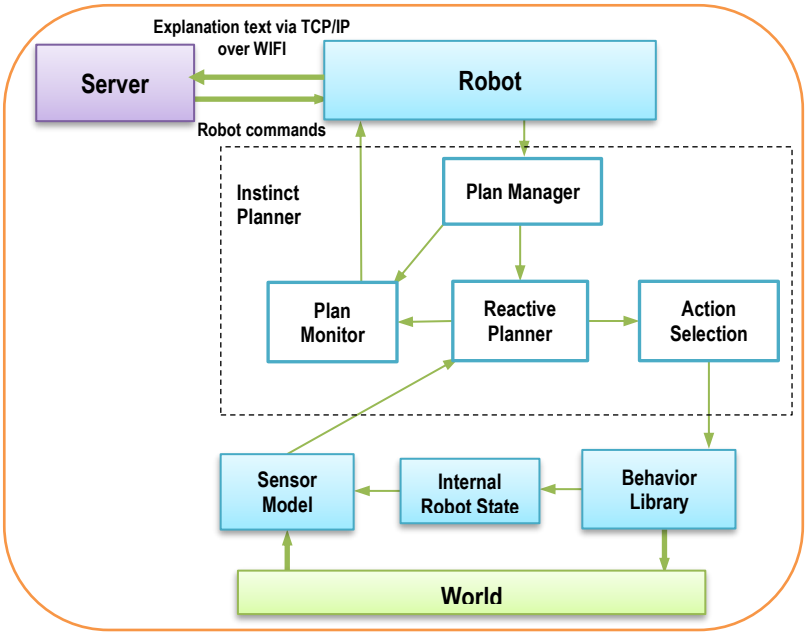

Fig. 8: Instinct reactive planner [80]

3.2.2 Automated Rationale Generation model - *Post-hoc domain-agnostic*

Automated rationale generation (ARG) is an explainable model proposed by Ehsan, et al. [1] for real-time explanation generation that learns to translate an autonomous agent's internal state and action data representations into natural language. It allows generating a natural language explanation for agent behavior as if a human had performed the behavior [71]. The intuition behind the rationale generation is that humans can engage in effective communication by verbalizing plausible motivations for their actions. The communication

can be effective even when the verbalized reasoning does not correlate actually with the decision-making neural processes of the human brain [137]. Ehsan, et al. [1] applied ARG for human-like explanation generation in the context of an agent that plays Frogger, a sequential environment, where decisions (i.e., selection of actions that maximize expected future rewards or utility) that the agent has made in the past influence future decisions. The approach is to collect a corpus of human-like explanation data (Fig. 3), a corpus of think-aloud data from players who explained their actions in a game environment and use the corpus to train a neural rationale generator (an encoder-decoder neural network) to enable agents to learn to generate plausible human-like explanations for their own behavior. While the results are promising, the potential limitations to the rationale generator may be the lack of a more grounded and technical explanations framework. The framework lacks interactivity that offers users the possibility to contest a rationale or to ask the agent to explain its decision in a different way. Another limitation may stem from the data collection process that introduces breakpoints to request explanation from the players. However, it is necessary to determine how to collect the necessary data in continuous-time and continuous-action environments without much interruption on the participants.

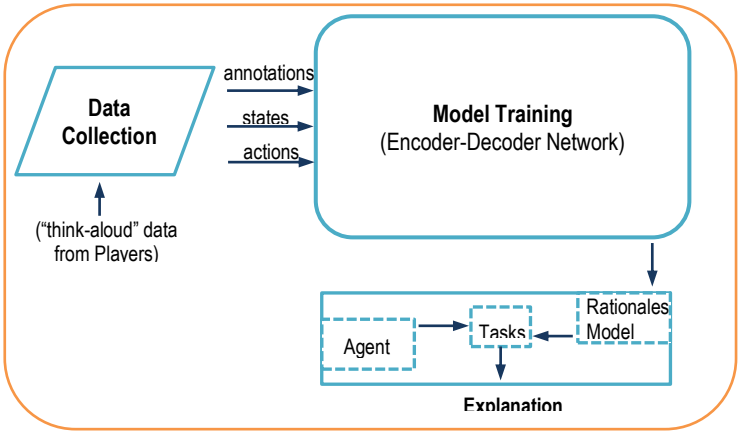

Fig. 9: Automated rationale generation [1]

3.2.3 **Autonomous Policy Explanation** – *Post-hoc domain-agnostic*

Autonomous Policy Explanation is a strategy proposed by Hayes and Shah [45] for a class of (reactive agent) robot controllers that rely on black-box trained reinforcement learning models [18], 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 aim is to enable the robot to explain its control policies, i.e., to reason over and answer questions about its underlying control logic, independent of its internal representation, enabling the human co-workers to synchronize their expectations (or mental models) and to identify faulty behaviour in the robot controller. The system learns a domain model (i.e., set of states) of its operating environment and the robot's underlying control logic (its policy) from real or simulated demonstrations or observations of the controller's execution traces (using a Markov Decision Process (MDP) framework as a basis for constructing the domain and policy models of the control software). The simulated observations are composed of annotations derived from the logging of function calls and their parameterizations alongside the current values of state variables at the run-time of the

controller. These are compiled into a single graphical model, capturing the important relational information between states and actions. For natural language communication, communicable predicates, i.e., Boolean classifiers similar to traditional STRIPS-style [138] planning predicates with associated natural language descriptions, are employed to convert attributes from sets of states into natural language and vice versa. The algorithms can then enable the agents to answer a direct inquiry on behaviour related questions on the environmental conditions (Fig. 11): for example, questions requesting explanation on occurrence, e.g., "When do you do __?", or to identify which robot behaviours will occur under a specified set of environmental conditions, e.g., "What do you do when __?", or an explanation for why a particular behaviour did not occur, e.g., ("Why didn't you do __?").

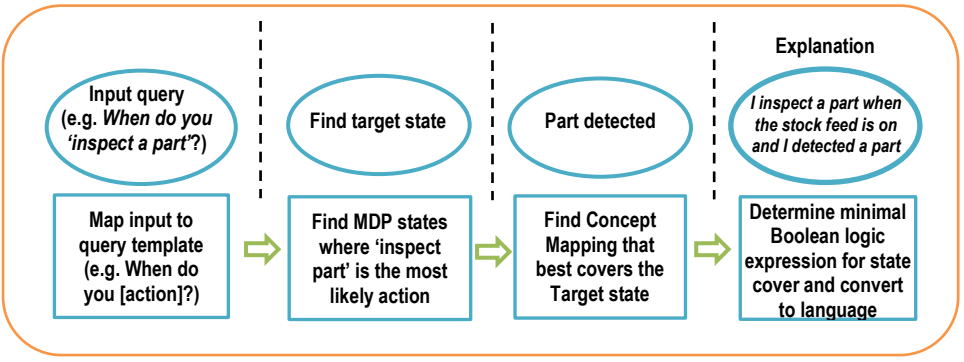

Fig. 10: Automated policy generation [45].

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

Explainable reinforcement learning (XRL) is a technique of explainability proposed in several studies for a class of reactive reinforcement learning (RL) agents [29]. Reactive RL agents are mostly (model-free) agents that select their actions based solely on their current observations [139]. Typically, they rely on a simple behavioral (state-action) policy scheme that enables them to learn a policy, i.e., a mapping from states to actions, given trial-and-error interactions between the agent and the environment [140]. RL agents do not need to plan or reason about their future to select actions, which obviously makes it hard for them to explain their behaviors. An RL agent would know at the end of each learning objective that choosing one action is preferable over others, or that some actions are associated with a higher value to attain the goal—but not why that is so or how it came to be. The "why" behind decision-making is lost during the agent's learning process as the policy converges to an optimal action-selection mechanism. 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 [29], Minimal Sufficient Explanation (MSX) via Reward Decomposition[30], and RARE [31].

Memory-based eXplainable Reinforcement Learning (MXRL) is an explainable reinforcement learning (XRL) strategy proposed by Cruz, et al. [29] to enable an RL agent to explain its decisions by using the probability of success, the number of transitions to reach the goal state, and an added episodic memory. Once it determines the probability of reaching the final state, the agent can provide the end-user a more comprehensible explanation for why one action was preferred over others. The RL agent can explain its behaviour not only in terms of Q-values or the probability of selecting an action but in terms of the necessity of

completing the intended task. Consequently, by accessing the memory, the agent's behaviour could be understood based on its experience by introspection or making some key analysis such as environment analysis - to observe certain and uncertain transitions, interaction analysis - to observe state-action frequencies, and meta-analysis - to obtain combined information from episodes and agents [140]. MXRL is implemented in a simulated scenario, a bounded grid world and an unbounded grid world with aversive regions. Using information extracted from the memory, the RL agent is shown to be able to explain its behaviour in an understandable manner for non-expert end-users at any moment during its operation. MXRL, however, it suffers some limitations with regards to the use of memory in large solution spaces.

Minimal Sufficient Explanation (MSX) is an XRL strategy proposed by Juozapaitis, et al. [30] to enable an explanation of the decisions of RL agents via reward decomposition. The approach is to decompose rewards into sums of semantically meaningful reward types so that actions can be compared in terms of trade-offs among the types. MSX is expected to provide a compact explanation, in a domain-independent framework, of why one action is preferred over another in terms of the reward types. It exploits an off-policy variant of Q-learning that converges to an optimal policy and the correct decomposed action values. The focus is on explanations that learn Q-functions that allow for observing how much an agent prefers one action over another. 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 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 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 (by Tabrez and Hayes [31]) to address the need for establishing a shared behaviour (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 modelled upon the assumption that sub-optimal collaborator behaviour is the result of a misinformed understanding of the task rather than a problem with the collaborator's rationality. Thus, using a Markov Decision Process, the human's sub-optimal decision-making is attributable to a malformed policy given an incorrect task model. By these assumptions, RARE can infer the most likely reward function used as a basis for a human's behaviour through interactive learning and reward update; identify the single most detrimental missing piece of the reward function; and then communicate this back to the human as actionable information to enable the collaborator to update their reward function (task comprehension) and policy (behavior) while performing the task and not after the task is completed. This process should enable the robot to provide a human with a policy update based on perceived model disparity, reducing the likelihood of costly or dangerous failures during joint task execution. RARE is implemented in a color-based collaborative Sudoku variant and an autonomous robot (Rethink Robotics Sawyer) [31]. The robot is reported to interrupt users that are about to make mistakes, informing that such action will cause task failure, and explaining which game constraint will inevitably be violated. However, the RARE model still lacks the comprehensibility of its policy.

Other XRL for model-free RL agents include the work of Madumal, et al. [141] that utilizes causal models to generate contrastive explanations (e.g., "why" and "why not" questions)

as a way to explain agent behaviour in a partially observable 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. Another work by Pocius, et al. [142] utilized saliency maps, also, to explain agent decisions in a partially observable game scenario; focusing mainly on deep RL to provide visual explanations. However, saliency maps did not help to explain long term causality and can be sensitive. The study by Sequeira and Gervasio [140] provided explanations through introspective analysis of the RL agent's interaction history. The framework analyzes an agent's history of interaction with the environment to extract interestingness elements that help explain its behavior.

3.3 Hybrid XGDAI

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

PeCoX is a framework proposed by Neerincx, et al. [59] for the development of explanations from an agent's perceptual and cognitive foundations (Fig. 12). PeCoX's perceptual level entails the use of an Intuitive Confidence Measure (ICM) and the identification of a counterfactual reference. The cognitive level entails the selection of beliefs, goals and emotions for explanations.

PeCoX's perceptual ICM is model-agnostic, designed for any machine learning model as it depends solely on the input and output of a trained model and future feedback about that output. The confidence (or uncertainty) reflects the machine learning model's expected performance on a single decision or classification. ICM is designed based on the notion of similarity and previous experiences: previous experiences with the ML model's performance directly influence the confidence of a new output, which is based on how similar the past data points are to the new data point.

PeCoX's cognitive framework considers explanations from the intentional stance [143]. The notion of the intentional stance assumes that the action is a consequence of the intentions of the agent performing the action. Explanation of the agent's action is then provided by giving the reasons for the underlying intention. Such explanation typically consists of beliefs, goals, and/or emotions, e.g. 'I hope (emotion) that you will take my advice to eat vegetables every day because I want (goal) you to adopt a healthy lifestyle, and I think (belief) that you currently do not eat enough vegetables'[144, 145]. PeCoX's cognitive framework is domain-agnostic.

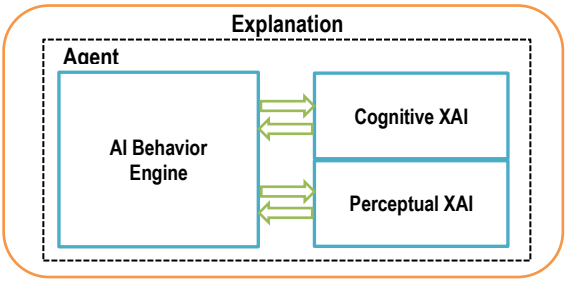

Fig. 11: PeCoX explanation generation framework[59]

Explainability Techniques	References	Deliberative/ Reactive/ Hybrid agents	Transparent/ Post-hoc	Domain-Specific/ Domain-Agnostic
Goal-Driven Autonomy (GDA)	[91], [146], [54]	D	T	DS
Explainable BDI model	[65], [123]	D	T	DS
SAT	[125]	D	T	DS
Meta-AQUA Explanation model	[126]	D	T	DS
Proactive Explanation	[106]	D		
EBG	[128], [38], [41], [129]	D	P	DS
KAGR	[49]	D	P	DA
eXplainable-PLAN	XAI-PLAN[51], **RBP** [132], **WHY-PLAN** [34]	D	P/T	DA
Explainable NPC	[96], [135]	D	P	DA
Debrief	[37]	D	T	DS
Explainable CREATIVE Model	[43], [136]	D	T	DS
Transparent Reactive Planning Technique	[80]	R	P	DA
Automated Rationale Generation (ARG)	[1]	R	P	DA
Autonomous Policy Explanation	[45]	R	P/T	DA
Explainable Reinforcement	[29], [139], [140], [30], [31]	R	P	DA

Learning (XRL)				
Perceptual-cognitive explanation (PeCoX)	[59], [143], [144, 145]	H	P	DA

Table 3. Summary of explainability Techniques for XGDAI

4 Explanation Communication Techniques for XGDAI

This section deals with the question of what exactly is communicated by the agent to the end-user, as an explanation, and how it is presented. This is characterized by the form of explanation (e.g., textual, visual, speech, etc.), and the content of the explanation (i.e., what is communicated by the agent) [59]. Essentially, this section distinguishes explanation communication for XGDAI according to verbal (e.g., speech) and non-verbal communication (visualization, expressive light – state displays, expressive motion – gestures, logs, and text). The following sections clarify this point.

4.1 Visualization techniques

Visualization is a technique of communicating agents' plans by externalizing the various pathways involved in the decision support process. This technique builds visual mediums between the planner and the humans to establish trust and transparency between the humans and the machine. An important justification for plan visualization is the need to minimize the time taken to communicate the agents' plans in natural language to the humans in the loop. Chakraborti, et al. [47] proposed a visualization approach to explainable AI planning for Mr. Jones (an end-to-end planning agent) to externalize the "mind" of the agent – i.e., the various processes that feed the different capabilities of the agent. Mr. Jones uses a set of widgets that give users a peek into its internal components. One widget presents a word cloud representation of Mr. Jones's belief for a given task. Another widget shows the agents that are in Mr. Jones's environment captured using four independent camera feeds - which helps the agent to determine what kind of task is more likely. This information is obtained via snapshots (sampled at 10-20 Hz) presented in a third widget. Finally, a fourth widget represents a word cloud-based summarization of the audio transcript of the environment. This transcript provides a brief representation of the things that have been said in the agent's environment in the recent past via the audio channels.

4.1.1 Some state-of-art examples of visualization techniques for agent perceptual function

Class Activation Maps

Deep learning architectures based on convolutional neural networks (CNN) achieve state-of-the-art results in many computer vision tasks. From the training data, CNNs learn to extract a deep hierarchy of task-relevant visual features. 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 classification of an image. Class Activation Maps (CAM) (1) 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, it can also be visualized what regions of the input image are relevant for the distinction. Figure 12 shows an example of a CAM from a ResNet50 trained on ImageNet for the category “Egyptian cat”.

Fig. 12: Class Activation Mapping for the classification “Egyptian cat” using RestNet50 trained on ImageNet in a Keras (4) implementation. The heat map highlights the discriminative image regions that contributed to ResNet’s classification. (Image from Wiki Commons, “cat” by xmhuqijian@yahoo.cn / CC BY)

Class Activation Maps, as introduced by Zhou et al. (1), work by inserting a global average pooling (GAP) layer directly after the last convolutional layer of the architecture. 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. To address this issue, Selvaraju et al. (2) introduced Gradient-weighted Class Activation Mappings (Grad-CAM) based on gradients flowing into the final convolutional layer. 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. (3) use a three-dimensional CNN to analyze MRI data of possible migraine patients; they use CAMs to highlight the discriminative brain areas. In such applications, CAMs can be used to guide a medical expert's attention to different image regions and to assess the network's prediction.

4.2 Expressive light

Expressive light is an approach to explanation communication that explores the use of lights for visualization of the robot’s internal state in relation to both tasks and its environment. The technique offers a large degree of choices in terms of animation pattern, color, and speed. A useful justification for expressive light is for communication 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. [86] explored the use of lights as a medium of communication on a mobile robot called CoBot (Fig. 12). The robot interacts with humans in different specific manners: requests help (to activate objects in its tasks), influences change in the user’s motion (in relation to its own motion) or provides useful information (task-related or general). To communicate its internal state, a node (node 1) running on the robot collects state information at every time

step. Whenever a change in state occurs, this change triggers a command to the microcontroller (node 2), notifying it only of the variables in the state that changed. The microcontroller is programmed with a state-animation mapping algorithm that triggers the animation in the programmable lights (3) upon notification of each state change.
In a related work, Song and Yamada [83] explored the use of expressive light on a Roomba robot (an appearance-constrained robot) to study people's perception and interpretation of the robot. Using two light expressions, namely, green in a low-intensity (GL) and red in a high-intensity (RH) as a way to communicate, people could construct rich and complex interpretations of the robot's behavior, although such interpretations are heavily biased by the design of expressive lights.

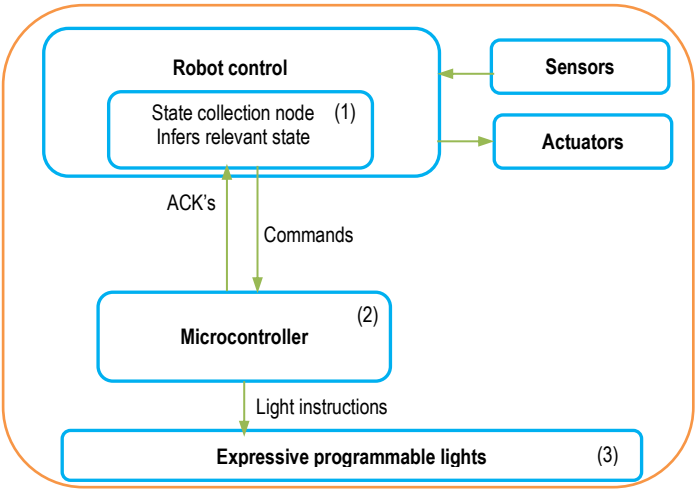

Fig. 13: CoBot's Control diagram of proposed expressive lights interface [86].

4.3 Expressive motion

For robots wandering among pedestrians in public space (e.g., service robot, etc.), one of the concerns is that it is difficult for people to understand their intentions or meanings behind their actions (e.g., motions or movements, etc.), for example, which direction the robot is going to take, or what the robot is going to do. People's understandability of such robots is also particularly influenced by the robot's appearance. The motivation here is thus to make the robot behave more like humans who often employ non-verbal expressions such as changes in facial and bodily expression. In this context, expressive motion is suggested as a useful technique to communicate robots' intentions to pedestrians to improve their mental impressions of the robot. As an example, Mikawa, et al. [84] proposed the use of rotational head movement for a teleoperated mobile robot as a way of expressing its intent to pedestrians in public spaces. The robot can express the intention by rotating its head to look where it is going whenever it changes its traveling course around pedestrians. A human operator teleoperates the robot for safety, setting a target position and moving speed and direction for the robot. Consequently, the direction of the robot head is determined by an artificial potential field (APF) generated based on the specified goal position as well as the positions of both pedestrians and obstacles around the robot. In effect, surrounding people can know the robot's intention of trying to change a traveling direction in advance. Other related work on expressive motion using gestures and eye gaze can be found in the review of [147]

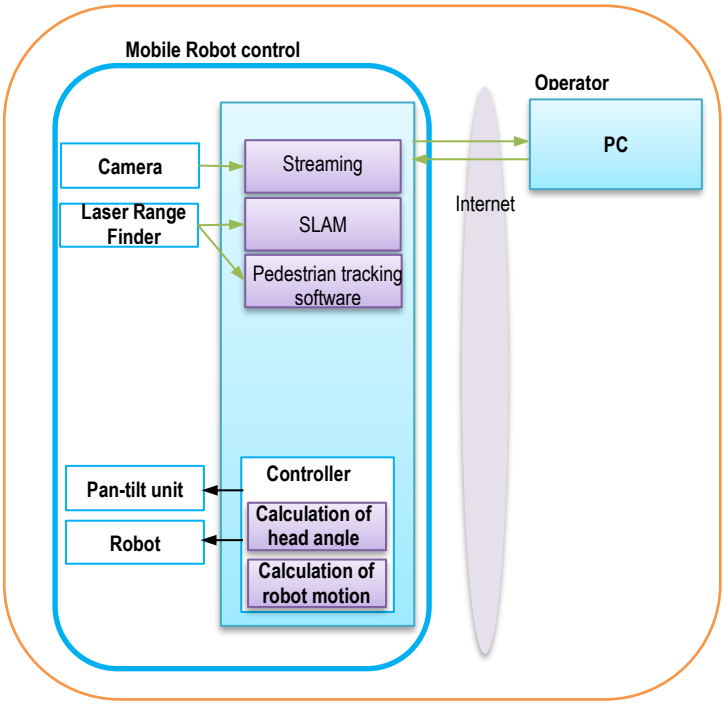

Fig. 14: Configuration of mobile robot teleoperation system for expressive head motion [84].

4.4 Logs

For agents that rely on relational (or symbolic) logics for knowledge representation and reasoning, the use of logs of data comes easily as a means of communicating the internal state of the robots. An example is the explainable BDI model proposed by Harbers, et al. [65], which stores all past mental states and actions of the agent needed for explanations in a behaviour log. When there is a request for an explanation, the explanation algorithm is applied to the log selecting the beliefs and goals that become part of the explanation. Other examples are the eXplainable AI (XAI) architecture proposed by Van Lent, et al. [135] for NPCs in Full Spectrum Command which works during the after-action review phase to extract key events and decision points from the playback log; and CREATIVE proposed by Wicaksono and Sheh [43] which stores all relevant information needed for explanations, including the learned hypotheses, in Prolog. CREATIVE utilises relational representation, so its results are inherently explainable as it describes the relation between objects as Prolog facts.

4.5 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 on a daily basis, they require skills for natural conversational ability. The earliest generation of humanoid robots (i.e., developed in the 1970s and 1980s) in this category were essentially equipped with such conversational skills, although the skills were mostly primitive at the time. They were typically designed as simple combinations of speech input/output mappings[148]. A good example is the Waseda Robot, WABOT-1 [149], 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)[150]. The WABOT-1 system could accept Japanese spoken command sentences, only in the form of primitive strings of separately spoken Japanese words, and then

respond to the meaning of the command in speech to make the robot move as commanded. The inner core of the system works as an automaton, makes transition after the recognition of an input sentence, and simultaneously makes an output. A further upgrade on the conversational system to make the speech more natural, particularly the speech synthesis part, was introduced in WABOT-2 [151], 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 the words, vowel durations of each unit and the accent patterns.

Another more recent social robot in this class is PaPeRo [152], a childcare robot that is designed with the capability to converse with humans in a more natural way. 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 with a microphone for human-robot communication, which is equipped with a near-field direction of arrival estimation, noise cancellation, and echo cancellation [153]. PaPeRo could 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 [154] which uses a commercial hardware electronic system for speech synthesis and ASKA[155] receptionist robots, with conversational speech dialogue system that can recognize user's question utterances, and answer the user's question by a text-to-speech voice processing with a hand gesture and head movement. Robovie [156] is another example in this category that is capable of conversing in English using a vocabulary of about 300 sentences for speaking and 50 words for recognition. The reader is referred to the work of Leite, et al. [157] for a more comprehensive review on social robots.

5 Continual Learning for Explainability

In this section, we examine techniques in XGDAI that enable *continual* learning of domain knowledge, domain model, 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 explainability in deliberative 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 hand-crafted 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. CBR is one of the techniques adopted in many works on XGDAI to enable *continual* learning of domain knowledge to minimize the amount of domain engineering for explanation generation. CBR is a learning and adaption technique that can help to build an explanation by retrieving and adapting past experiences, which are stored in a case-base or case-library in the form of cases [158]. Learning, in the CBR paradigm, implies extending the agent's knowledge by interpreting new experiences and incorporating them into memory (i.e., the case-library) or by re-interpreting and re-indexing old experiences to make them more usable and accessible [159]. It also implies the formation of generalizations over a set of experiences. Interpreting an experience in CBR means creating an explanation that connects the agent's goals and actions with resulting outcomes. There

are two main learning approaches in CBR: learning by observation (supervised) and learning by own experience (unsupervised) [158]. Learning by observation [160] happens when the case-library is populated by direct observation of real data or from expert demonstration. Learning by own experience [161] is done after each reasoning cycle where the proposed solution is examined. If successful, then it can be stored and used for future reference [158]. A learning by demonstration (observation) approach was implemented by Weber, et al. [146] to reduce the amount of domain engineering necessary to implement the agent that plays the real-time strategy (RTS) game, StarCraft. The CBR was applied to learn expectations, explanations, and goals from expert demonstrations [146]. Using two case libraries: (1) an adversary library to provide examples for adversary actions and (2) a goal library to select goals for the agent to pursue, the learning system could generate explanations when the agent's active goal completes, or when a discrepancy is detected by applying a goal formulation to the adversary, using the adversary case library. In a similar effort, Floyd and Aha [54] applied two **case-based reasoning (CBR)** systems to assist agent learning and explanation generation. Both CBRs use cases that are learned while interacting with the operator (learning by observation). The first CBR process evaluates the robot's trustworthiness and selects a new behaviour if the robot is behaving in an untrustworthy manner. When the robot adapts its behaviour, a second case-based reasoning process is used to generate an explanation for why the change occurred. The agent's explanations are based on explicit feedback received from an operator. The model is evaluated on a simulation environment that involved an operator instructing a robot to patrol in the environment, to identify suspicious objects, and to classify them as threats or harmless. Consequently, the robot would evaluate its own trustworthiness and adapt its behaviour if it determined its behaviour was untrustworthy.

5.2 Explainable Reinforcement Learning (XRL)

For many RL agent, one of the major concerns is a loss of information about the decision process during the agent's policy learning process. An RL agent would know at the end of each learning objective that choosing one action is preferable over others, or that some actions are associated with a higher value to attain the goal, but the "why" behind decision-making is lost during the agent's learning process as the policy converges to an optimal action-selection mechanism. The lack of bookkeeping, traceability, or recovery of this processed, once an optimal policy has been learned, makes it hard for the agent to explain itself or transfer learned policy for explainability. A few extensions to the application of the XRL technique seek to enhance retention of the learned policy for explaining agents' decision process. Some examples include the Memory-based eXplainable Reinforcement Learning (MXRL) proposed by Cruz, et al. [29] 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 make introspection, observation or analysis on its environment transitions and interactions. The RL agent is shown to be able to explain its behaviour or decision to non-expert end-users at any moment during its operation, 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 behavioural architectures for GDAI - e.g. deliberative, reactive, and hybrid - come

with different techniques for explanation generation. The state-of-the-art suggests the need for an effective unified approach towards explainabilty in XGDAI. Overall, many explainability techniques are still lacking 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 roadmap for effective actualization of explainability in XGDAI.

6.1 Road Map 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 explanation. In many such applications, agents already have a set of plans and clear decision system to achieve a goal thus constructing very rich explanation are often relatively straightforward. However, applying these techniques can be problematic as agents would perform optimal in a specific domain and suboptimal in other domains where knowledge of the agent's world are 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 enable cross-platform explanation generation for agents and robots. Majority of these techniques can be found for reactive black box agents whose decisions are based solely on current environmental state, not a priori defined. Without a model of the world or domain knowledge, these agents can generate explanation 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 benefits both deliberative and reactive agents.

6.1.1 Theory of Mind for Agent's Teammates

A significant body of work on XGDAI involve agents/robots collaborating with humans and other agents. Given this reality, it is therefore imperative for agents to adequately understand their teammates for effective collaboration and to provide useful and timely explanation when necessary. A useful step in this direction may be to integrate a theory of the mind (ToM) concept in the explanation framework enabling an agent to also reason about the perception and mental state of other teammates. A well-constructed ToM should enable the robot to understand the expectations of its teammates and thereby provide useful, relevant and timely explanation when necessary. The motivation here is that humans are well known to collaborate with their teammates and generate explanation for their behaviour using extensively the ToM concept. Theory of Mind is the ability to reason about other peoples' perception, beliefs and goals and to take them into account (ToM)[162].

6.1.2 Rich Perceptual-Cognitive Explainability framework

A significant number of literature (particularly literatures on deliberative XGDAIs) provide 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 agent's perceptual function (sensor information, environment states, etc.) with poorly or non-existent explainable cognitive framework or explainable decision-making framework. A rich perceptual-cognitive explainable framework that abstract low-level agent's perception (primarily perceptual explanatory knowledge) for high level cognition and explainability would significantly advance the current work on XGDAI.

6.1.3 Natural Language Processing

In the field of machine learning, natural language processing (NLP) is the ability of a computer system or AI system to understand, analyse, manipulate, and potentially generate human language [163]. For XGDAI, NLP is crucial in the explanation generation/communication framework to enable human comprehensible explanation of agent's/Robot's perception, cognition, or decisions. Currently, the state of art in XGDAI reveals less extensive or even non-existent natural language processing ability for many of the agent and robots surveyed. The addition of a rich NLP system to existing frameworks would significantly benefit XGDAI in terms of their usefulness and applicability.

6.1.4 Continual learning of explanatory knowledge

As agents/robots interact with their environment, teammates, and supervisors, they are expected to provide explanation for their decisions or actions in different situations and scenarios. Handcrafting of explanatory knowledge to satisfy all possible situations and expectations is difficult and would require significant effort or domain engineering to accomplish. In this respect, agent's ability to continuously learn to generate/construct explanation in different situations is therefore crucial to the success of explainable agents. As in traditional machine learning, learning in this case could be achieved by supervised learning (e.g. learning from demonstration [160]), unsupervised learning[161], and reinforcement learning. Currently, CBR[158] and XRL [29] techniques have been applied in a few studies on deliberative and reactive agents respectively to address this concern. For reactive XRL, however, a major concern is how to retain policy learnt by agent for constructing explanation. There is still a significant gap to fill up in this direction. Issues of scalability and resource management would also need to be addressed if explanatory knowledge are to be stored in a continual learning framework.

6.1.5 Integrating verbal and non-verbal explanation communication

The current state of art reveals different explanation communication modalities for XGDAI applied in separate niche areas/scenarios. With diverse application scenarios, the need to communicate agent's/robot's plan, decision, intentions, etc. by a combination of both verbal and nonverbal communication means is necessary if such explanation should be natural and effective. A relevant example is seen from how humans explain/communicate using both verbal (i.e. speech) and nonverbal (e.g. gesture, facial expression, emotion, etc.) means, depending on which is most effective for the circumstance. XGDAI should also enable such combination of different modalities for example, for agents sharing a pedestrian walkway with humans, rotational head/eye movements seems more likely natural and effective to communicate the agent's decision to turn left, change path, etc[]; while 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, the use of speech for normal explanatory conversation is necessary, however expressive motion like gesture (e.g. head nodding, etc.) would be useful to provide tacit explanation, or expressive light[] or sound to alert teammates of its mental state in emergency situations. There can be many

possible effective and natural combination of explanation communication modalities. More research work is still required to bridge this gap,

7 Conclusions and further work

The field of XGDAI is emerging with many rich applications in several domains. Explainability enables transparency for these types of agents/robots and encourages users' trust for applications in safety critical situations. These survey presents several techniques for explanation generation and communication proposed and implemented in XGDAI till date. Typically, many XGDAIs techniques can enable the robots/agents to provide justification/explanation for their decisions/plan and rationale for their actions. However, the state of art shows that current works done on XGDAIs are still in their infancy lacking an extensive explanation generation and communication framework. Consequently, this study highlights a roadmap actualization of an extensive XGDAI that has an extended perceptual and cognitive explanation capability.

Future work will involve the development of a transparent integrated architecture, domain agnostics, for effective actualization of explainability in XGDAI.

References

[1] 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: ACM, pp. 263-274.

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

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

[4] A. Bundy, "Preparing for the future of Artificial Intelligence," ed: Springer, 2017.

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

[6] 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*, 2019: International Foundation for Autonomous Agents and Multiagent Systems, pp. 1078-1088.

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

[8] O. Biran and C. Cotton, "Explanation and justification in machine learning: A survey," in *IJCAI-17 workshop on explainable AI (XAI)*, 2017, vol. 8, p. 1.

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

[10] 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, p. 93, 2019.

[11] A. Holzinger *et al.*, "Towards the augmented pathologist: Challenges of explainable-ai in digital pathology," *arXiv preprint arXiv:1712.06657*, 2017.
[12] P. Langley, B. Meadows, M. Sridharan, and D. Choi, "Explainable agency for intelligent autonomous systems," in *Twenty-Ninth IAAI Conference*, 2017.
[13] 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.
[14] J. Park, D. 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, 2018.
[15] M. Schmerling, G. Schillaci, and V. V. Hafner, "Goal-directed learning of hand-eye coordination in a humanoid robot," in *5th Joint International Conference on Development and Learning and Epigenetic Robotics, ICDL-EpiRob 2015*, 2015, pp. 168-175.
[16] A. D. Dragan, K. C. Lee, and S. S. Srinivasa, "Legibility and predictability of robot motion," in *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*, 2013: IEEE Press, pp. 301-308.
[17] D. Dannenhauer, M. W. Floyd, M. Molineaux, and D. W. Aha, "Learning from Exploration: Towards an Explainable Goal Reasoning Agent," 2018.
[18] 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*, 1998, pp. 483-488.
[19] 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)*, 1994, vol. 3: IEEE, pp. 1938-1945.
[20] Y. Wang, D. Mulvaney, I. Sillitoe, and E. Swere, "Robot navigation by waypoints," (in English), *J Intell Robot Syst*, vol. 52, no. 2, pp. 175-207, Jun 2008.
[21] P. Davidsson, *Autonomous agents and the concept of concepts*. Department of Computer Science, Lund University, 1996.
[22] T. Hellström and S. Bensch, "Understandable robots-what, why, and how," *Paladyn, Journal of Behavioral Robotics*, vol. 9, no. 1, pp. 110-123, 2018.
[23] 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: International Foundation for Autonomous Agents and Multiagent Systems, pp. 2180-2182.
[24] R. H. Wortham and A. Theodorou, "Robot transparency, trust and utility," *Connection Science*, vol. 29, no. 3, pp. 242-248, 2017.
[25] L. Takayama, D. Dooley, and W. Ju, "Expressing thought: improving robot readability with animation principles," in *2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2011: IEEE, pp. 69-76.
[26] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138-52160, 2018.
[27] S. Sreedharan and S. Kambhampati, "Balancing explicability and explanation in human-aware planning," in *2017 AAAI Fall Symposium Series*, 2017.

[28] K. Knight, "Are many reactive agents better than a few deliberative ones?," in *IJCAI*, 1993, vol. 93, pp. 432-437.
[29] F. Cruz, R. Dazeley, and P. Vamplew, "Memory-Based Explainable Reinforcement Learning," in *AI 2019: Advances in Artificial Intelligence*, Cham, 2019: Springer International Publishing, pp. 66-77.
[30] 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.
[31] 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)*, 2019: IEEE, pp. 751-753.
[32] M. Wooldridge, "Conceptualising and developing agents," in *Proceedings of the UNICOM Seminar on Agent Software*, 1995, vol. 42: London.
[33] A. L. Hayzelden and J. Bigham, *Software agents for future communication systems*. Springer Science & Business Media, 1999.
[34] R. Korpan and S. L. Epstein, "Toward Natural Explanations for a Robot's Navigation Plans.(2018)," 2018.
[35] A. S. Rao and M. P. Georgeff, "BDI agents: from theory to practice," in *ICMAS*, 1995, vol. 95, pp. 312-319.
[36] D. Chapman, "Planning for conjunctive goals," *Artificial intelligence*, vol. 32, no. 3, pp. 333-377, 1987.
[37] W. L. Johnson, "Agents that Learn to Explain Themselves," in *AAAI*, 1994, pp. 1257-1263.
[38] 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.
[39] R. J. Mooney and S. Bennett, "A Domain Independent Explanation-Based Generalizer," in *AAAI*, 1986, pp. 551-555.
[40] 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: Morgan Kaufmann Publishers Inc., pp. 968-973.
[41] J. W. Shavlik, "Acquiring Recursive Concepts with Explanation-Based Learning," in *IJCAI*, 1989, pp. 688-693.
[42] M. Sridharan and B. Meadows, "Towards a Theory of Explanations for Human–Robot Collaboration," *KI-Künstliche Intelligenz*, pp. 1-12, 2019.
[43] 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.
[44] M. R. Wick and W. B. Thompson, "Reconstructive Explanation: Explanation as Complex Problem Solving," in *IJCAI*, 1989, pp. 135-140.
[45] 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*, 2017: IEEE, pp. 303-312.
[46] 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)*, 2017: IEEE, pp. 676-682.
[47] T. Chakraborti *et al.*, "Visualizations for an explainable planning agent," *arXiv preprint arXiv:1709.04517*, 2017.

[48] 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 …, 2014.

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

[50] 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)*, 2013: IEEE, pp. 1-6.

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

[52] 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)*, 2017: IEEE, pp. 3427-3432.

[53] 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)*, 2017: IEEE, pp. 88-93.

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

[55] 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.

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

[57] M. Lomas, R. Chevalier, E. V. Cross II, 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: ACM, pp. 187-188.

[58] 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*, 2009: Citeseer.

[59] 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*, 2018: Springer, pp. 204-214.

[60] 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)*, 2018: IEEE, pp. 1005-1011.

[61] 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: International Foundation for Autonomous Agents and Multiagent Systems, pp. 997-1005.

[62] 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), 2015: IEEE, pp. 7-12.

[63] 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*, 2015: Springer, pp. 224-241.

[64] R. Sheh and I. Monteath, "Introspectively Assessing Failures through Explainable Artificial Intelligence," in *IROS Workshop on Introspective Methods for Reliable Autonomy*, 2017.

[65] 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*, 2010, vol. 2: IEEE, pp. 125-132.

[66] 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*, 2010: Springer, pp. 28-39.

[67] 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.

[68] 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*, 2009: Springer, pp. 168-182.

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

[70] N. C. Codella *et al.*, "Teaching meaningful explanations," *arXiv preprint arXiv:1805.11648*, 2018.

[71] 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: ACM, pp. 81-87.

[72] E. Groshev, A. Tamar, M. Goldstein, S. Srivastava, and P. Abbeel, "Learning generalized reactive policies using deep neural networks," in *2018 AAAI Spring Symposium Series*, 2018.

[73] M. Guzdial, J. Reno, J. Chen, G. Smith, and M. Riedl, "Explainable PCGML via Game Design Patterns," *arXiv preprint arXiv:1809.09419*, 2018.

[74] S. Rosenthal, S. P. Selvaraj, and M. M. Veloso, "Verbalization: Narration of Autonomous Robot Experience," in *IJCAI*, 2016, pp. 862-868.

[75] M. K. Sahota, "Reactive deliberation: An architecture for real-time intelligent control in dynamic environments."

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

[77] F. Wang *et al.*, "Residual attention network for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3156-3164.

[78] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048-2057.

[79] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4651-4659.

[80] 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.

[81] A. D. Dragan, S. Bauman, J. Forlizzi, and S. S. Srinivasa, "Effects of robot motion on human-robot collaboration," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, 2015: ACM, pp. 51-58.

[82] 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: ACM, pp. 247-248.

[83] 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: ACM, p. LBW629.

[84] 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)*, 2018: IEEE, pp. 249-254.

[85] T. Chakraborti, S. Sreedharan, Y. Zhang, and S. Kambhampati, "Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. IJCAI International Joint Conference on Artificial Intelligence (2017), 156–163," ed, 2017.

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

[87] 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)*, 2015: IEEE, pp. 1-6.

[88] 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.

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

[90] M. Klenk, M. Molineaux, and D. W. Aha, "Goal-Driven Autonomy For Responding To Unexpected Events In Strategy Simulations," *Comput Intell-Us*, vol. 29, no. 2, pp. 187-206, 2013.

[91] M. Molineaux, M. Klenk, and D. Aha, "Goal-driven autonomy in a Navy strategy simulation," in *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

[92] 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.

[93] 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*, 2005: IEEE, pp. 708-713.

[94] 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*, 2006: IEEE, pp. 80-85.

[95] 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)," *IEEE Transactions on Cognitive and Developmental Systems*, Article vol. 8, no. 3, pp. 214-231, 2016, Art no. 7470616.

[96] 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.

[97] J. Vermeulen, "Improving intelligibility and control in ubicomp," 2010: ACM.

[98] 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.

[99] 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*, 2013: IEEE, pp. 3-10.

[100] 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)*, 2017: IEEE, pp. 1424-1431.

[101] 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: International Foundation for Autonomous Agents and Multiagent Systems, pp. 1495-1503.

[102] 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 2018 on International Conference on Multimodal Interaction*, 2018: ACM, pp. 557-558.

[103] J. Haspiel *et al.*, "Explanations and expectations: Trust building in automated vehicles," in *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018: ACM, pp. 119-120.

[104] 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.

[105] 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: ACM, pp. 179-180.

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

[107] 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*, 2018.

[108] M. R. Penner and S. J. Y. Mizumori, "Neural systems analysis of decision making during goal-directed navigation," *Progress in Neurobiology*, vol. 96, no. 1, pp. 96-135, 2012/01/01/ 2012.
[109] A. Jauffret, N. Cuperlier, P. Tarroux, and P. Gaussier, "From self-assessment to frustration, a small step toward autonomy in robotic navigation," (in English), *Front Neurorobotics*, vol. 7, 2013.
[110] R. Sukkerd, R. Simmons, and D. Garlan, "Towards explainable multi-objective probabilistic planning," in *Proceedings of the 4th International Workshop on Software Engineering for Smart Cyber-Physical Systems*, 2018: ACM, pp. 19-25.
[111] S. Penney, J. Dodge, C. Hilderbrand, A. Anderson, L. Simpson, and M. Burnett, "Toward foraging for understanding of StarCraft agents: An empirical study," in *23rd International Conference on Intelligent User Interfaces*, 2018: ACM, pp. 225-237.
[112] D. Amir and O. Amir, "Highlights: Summarizing agent behavior to people," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 2018: International Foundation for Autonomous Agents and Multiagent Systems, pp. 1168-1176.
[113] A. Grea, L. Matignon, and S. Aknine, "How explainable plans can make planning faster," 2018.
[114] 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: ACM, pp. 157-166.
[115] S. Stumpf, W.-K. Wong, M. Burnett, and T. Kulesza, "Making intelligent systems understandable and controllable by end users," 2010.
[116] 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.
[117] 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: ACM, pp. 181-186.
[118] T. Kulesza *et al.*, "Explanatory debugging: Supporting end-user debugging of machine-learned programs," in *2010 IEEE Symposium on Visual Languages and Human-Centric Computing*, 2010: IEEE, pp. 41-48.
[119] 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, 2009.
[120] R. Reiter and J. De Kleer, "An assumption-based truth-maintenance system," in *Artificial Intelligence*, 1986: Citeseer.
[121] F. C. Keil, "Explanation and understanding," *Annu. Rev. Psychol.*, vol. 57, pp. 227-254, 2006.
[122] B. F. Malle, "How people explain behavior: A new theoretical framework," *Personality and social psychology review*, vol. 3, no. 1, pp. 23-48, 1999.
[123] R. Flin and K. Arbuthnot, *Incident command: Tales from the hot seat*. Routledge, 2017.
[124] 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.

[125] M. R. Endsley, "Innovative model for situation awareness in dynamic defense systems," *Defense Innovation Handbook: Guidelines, Strategies, and Techniques*, 2018.
[126] M. T. Cox, "Perpetual self-aware cognitive agents," *AI magazine*, vol. 28, no. 1, pp. 32-32, 2007.
[127] M. T. Cox and A. Ram, "Introspective multistrategy learning: On the construction of learning strategies," *Artificial Intelligence*, vol. 112, no. 1-2, pp. 1-55, 1999.
[128] T. M. Mitchell, R. M. Keller, and S. T. Kedar-Cabelli, "Explanation-Based Generalization: A Unifying View," *Machine Learning*, vol. 1, no. 1, pp. 47-80, 1986/03/01 1986.
[129] J. W. Shavlik and G. F. DeJong, "BAGGER: An EBL system that extends and generalizes explanations," *Coordinated Science Laboratory Report no. UILU-ENG-87-2223*, 1987.
[130] 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.
[131] 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 *Twenty-Second International Conference on Automated Planning and Scheduling*, 2012.
[132] J. Bidot, S. Biundo, T. Heinroth, W. Minker, F. Nothdurft, and B. Schattenberg, "Verbal Plan Explanations for Hybrid Planning," in *MKWI*, 2010: Citeseer, pp. 2309-2320.
[133] S. Biundo and B. Schattenberg, "From abstract crisis to concrete relief—a preliminary report on combining state abstraction and htn planning," in *Sixth European Conference on Planning*, 2014.
[134] 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*, 2015: Springer, pp. 403-425.
[135] 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*, 2004: Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, pp. 900-907.
[136] N. Lavrac and S. Dzeroski, "Inductive Logic Programming," in *WLP*, 1994: Springer, pp. 146-160.
[137] N. Block, "Two neural correlates of consciousness," *Trends in cognitive sciences*, vol. 9, no. 2, pp. 46-52, 2005.
[138] R. E. Fikes and N. Nilsson, "A new approach to the application of theorem proving to problem solving," *Artificial Intelligence*, vol. 2, 1971.
[139] 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*, 1994, vol. 3: Cambridge, MA, p. 238.
[140] P. Sequeira and M. Gervasio, "Interestingness Elements for Explainable Reinforcement Learning: Understanding Agents' Capabilities and Limitations," *arXiv preprint arXiv:1912.09007*, 2019.
[141] P. Madumal, T. Miller, L. Sonenberg, and F. Vetere, "Explainable reinforcement learning through a causal lens," *arXiv preprint arXiv:1905.10958*, 2019.

[142] R. Pocius, L. Neal, and A. Fern, "Strategic tasks for explainable reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 10007-10008.
[143] D. C. Dennett, "Three kinds of intentional psychology," *Perspectives in the philosophy of language: A concise anthology*, pp. 163-186, 1978.
[144] M. Harbers, J. Broekens, K. Van Den Bosch, and J.-J. Meyer, "Guidelines for developing explainable cognitive models," in *Proceedings of ICCM*, 2010: Citeseer, pp. 85-90.
[145] S. A. Döring, "Explaining action by emotion," *The Philosophical Quarterly*, vol. 53, no. 211, pp. 214-230, 2003.
[146] B. G. Weber, M. Mateas, and A. Jhala, "Learning from demonstration for goal-driven autonomy," in *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.
[147] 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.
[148] C. Breazeal, A. Takanishi, and T. Kobayashi, "Social Robots that Interact with People," in *Springer Handbook of Robotics*, B. Siciliano and O. Khatib, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 1349-1369.
[149] K. Shirai, H. Fujisawa, and Y. Sakai, "Ear and Voice of the Wabot," *Bull. Sci. & Eng. Research Lab. Waseda Univ*, no. 62, 1973.
[150] 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*, 1974.
[151] T. Kobayashi, "Speech conversation system of the musician robot," *Proc. ICAR'85*, pp. 483-488, 1985.
[152] 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.
[153] M. Sato, A. Sugiyama, and S. i. Ohnaka, "Auditory system in a personal robot, PaPeRo," in *2006 Digest of Technical Papers International Conference on Consumer Electronics*, 2006: IEEE, pp. 19-20.
[154] 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*, 2002, vol. 3: IEEE, pp. 2478-2483.
[155] 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*, 2002, vol. 2: IEEE, pp. 1314-1319.
[156] 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)*, 2011: IEEE, pp. 37-44.
[157] I. Leite, C. Martinho, and A. Paiva, "Social robots for long-term interaction: a survey," *Int J Soc Robot*, vol. 5, no. 2, pp. 291-308, 2013.
[158] 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," *Auton Robot*, vol. 21, no. 1, pp. 65-78, 2006.
[159] J. Kolodner, *Case-based reasoning*. Morgan Kaufmann, 2014.

[160] M. van Lent and J. Laird, "Learning by observation in a complex domain," in *Proceedings of the Knowledge Acquisition Workshop*, 1998.
[161] R. C. Schank, *Dynamic memory: A theory of reminding and learning in computers and people*. cambridge university press, 1983.