Master in Operational Research

# Explainable Goal-driven AI
# A comprehensive review


*Student:*
Willy LAO

*Master's supervisor:*
Marcel MONGEAU

*ISAE-Supaero supervisor:*
Alain HAÏT

*Internship supervisor:*
Guillaume GAUDRON


Date: 05/06/2020

# Abstract

In the last recent years, Artificial Intelligence (AI) systems received an ever-increasing attention with the development of self-driving cars, service robots, recommendation agents or E-health systems. However, such safely-critical situations exacerbated the need of trust and transparency that black boxes algorithm such as Deep Neural Network cannot provide because of their opaque decision-making.

The majority of the studies about Explainable Artificial Intelligence concern data-driven systems. However, with the appearance of agents and systems like the one mentioned previously, research focusing on Goal-driven AI is still lacking.

This paper selected 130 recent articles in an exhaustive attempt to give a global literature review on Explainable Goal-driven agents and robots. First, it clarifies the different terminology used in Explainable AI research, then raises research questions and present how the selected papers answered them. Finally, it suggests different axis to address the future challenges and urgent needs that such systems bring.


# 1 What is Explainable Goal-driven AI

## 1.1 Explainable Goal-driven AI (XGDAI) vs Explainable Data-driven AI (XAI)

**Explainable AI** refers to techniques which enable systems and agents to provide justifications for their decisions, understandable by humans. It is an implementation of the social right to explanation.

As machine learning is booming and generates huge amounts of wealth, most of the works about transparency concern **Explainable Data-driven AI**. Some of its purposes is to find which features of the input data led to the decision, to understand the inner workings of the layers of a neural network in the case of deep learning, or to predict in a reproducible way the output given a certain data and context.

Even though autonomous agents and robots are have begun into our daily life, **Explainable Goal-driven AI** is a less researched branch of Explainable AI. It focuses on increasing the trust and understandability between robots and humans by studying their interactions, tapping into human psychology and looking for the best ways to present the agent's decisions. The user can identify the capabilities and limits of the agent and make an informed decision after figuring out the agent's inner workings.

Some of the studies make reference do the **Theory of Mind (ToM)**. It studies the ability to attribute mental states (beliefs, intents, knowledge, perspectives, etc.) to others and recognize that these mental states may differ from one's own. It is particularly important in human-robot collaboration where humans need to know the next action of the robot or understand its intention, expressed by visuals, speech, motions or other means.

## 1.2 Terminology clarification

In this section the attributes of AI agents are defined, their similarities and differences are highlighted. They are interchangeable between Data-driven AI and Goal-driven AI.

**Explainability**: An agent/robot is explainable if the decision it takes can be described with a certain logic and can be understood by humans. The explanation is a means for the intelligent agent and the human to dialogue.

**Transparency**: An agent/robot is transparent if it is possible to describe, inspect and reproduce the mechanisms through which an it makes decisions and learns to adapt to its environment.

**Understandability**: An agent/robot is understandable if a human is able to comprehend how it works without any need to explain its internal structure.

**Explicability**: A system is explicable if it can generate plans that are similar to the ones a human would expect, thus avoiding the need to provide justifications.

**Predictability**: Similar to Explicability, an agent/robot is predictable if its behavior matches the user's expectations.

**Legibility**: A robot is legible if an observer can infer its intention through its behavior. In goal-driven AI, the observer should be able to quickly know the goal of the robot.

**Readability**: A robot is readable if a human user can understand what it is doing and can predict its next action.

**Explicit explainability**: An agent/robot is explicitly explainable if it provides a direct and clear explanation of its decisions.

**Implicit explainability**: An agent/robot is implicitly explainable if its behavior is readable, legible, predictable, explicable and/or transparent enough, so that it can be inferred without needing to provide explicit explanations.

### 1.3 Explainable Goal-driven AI mechanisms

In order to provide easily understandable explanations to humans, Goal-driven AI can be divided into three mechanisms.

**Explanation Generation** is a phase that studies the inner model or the inner reasoning of an agent/robot. An explanation generation module is either added to the model or directly implemented in the decision loop to provide justifications of a result or of a behavior.

**Explanation communication** decides what elements of a justification have to be given. It also presents the explanation to the user or another agent in a certain way. It can be through text, visuals, log, speech, class activation maps, multi-modal interfaces or expressive motions and expressive lights for robots.

**Explanation reception** concerns how the user/observer understands the State of Mind (beliefs, intents, knowledge, perspectives…) of the AI agent/robot. Explanation efficiency is generally assessed with empirical evaluations, polls and social and human psychology studies.

## 2 Review methodology

### 2.1 Methodology

In this paper, we tried to have the most exhaustive picture of the current state of the Explainable Goal-driven AI. For this, 130 articles were selected using the following methodology.

<u>Sources:</u> Most of these different articles can be consulted in arXiv, aaai, Link Springer, IEEExplore, IJCAI or Researchgate.

<u>Selection criteria:</u>

**Recent papers (2007-2020)**: As the 2020 was not yet finished when this paper was released in June 2020, there are only 2 papers from 2020.

**Relevance for Explainable Artificial Intelligence**: For example, papers addressing explanations in social science without any relevancy to AI are excluded.

**Primary Study**: Only papers providing a direct contribution on Explainable Artificial Goal-driven AI (models, techniques, or explication interfaces) are included. Secondary studies like surveys are not.

**Explainable Agency**: Data-driven XAI research is not included. However, goal-driven agents/robots using Machine Learning mechanisms such as Reinforcement Learning were selected.

**Singularity**: Papers which have been published in an extended or complete version are not included.

## 2.2 Research questions

In this paper, several research questions are raised and figures will show how the selected papers studied them.

RQ 1: **Definition of XGDAI** – What is Explainable Goal-driven AI?

RQ 2: **Demographics** – How has the research on Explainable Goal-Driven AI evolved in the recent years?

RQ 3: **Subject** - What is the subject of Explainable Goal-driven AI?

RQ 4: **Recipient** – Who is the recipient of Explainable Goal-driven AI

RQ 5: **Applications scenarii** – What kinds of application scenarii are addressed or simulated in Explainable Goal-Driven AI studies?

RQ 6: **Motivations and needs** – What motivations and needs drive Explainable Goal-driven AI?

RQ 7: **Social Science and psychological background** – To which extent do studies about Explainable Goal-driven AI draw on social science and psychological background?

RQ 8: **Types of explanation** – What are the different types of explanations provided by Explainable Goal-driven AI?

RQ 9: **Granularity** – To which level of explanation should studies address Explainable Goal-driven AI?

RQ 10: **Techniques** – What kind of techniques, platforms or architecture are used in Explainable Goal-driven AI research?

RQ 11: **Presentation** – How are explanations presented for an efficient communication and reception in Explainable Goal-driven AI?

RQ 12: **Evaluation** – How are the explanations validated and evaluated in Explainable Goal-driven AI?

RQ 13: **Future work and challenges** – What future challenges Explainable Goal-driven AI has to strive for?

## RQ 1: Definition of XGDAI

Explainable Goal-driven AI (XGDAI) was already defined in sections 1.1 and 1.3. It can be divided into three main mechanisms: explanation generation, explanation communication and explanation reception.

## RQ 2: Demographics

The number of published studies are shown by year and by country in the figures 1 and 2. As the year 2020 is not yet finished at our paper's time of publication, only 2 papers were found in 2020.

With a maximum of 8 papers published in a year among the selected works, Explainable Goal-driven AI gathered little attention before 2015. At that time, research about it was mostly the investigation of a single laboratory like the

5 Dutch studies in 2010 or the 4 French studies in 2014.

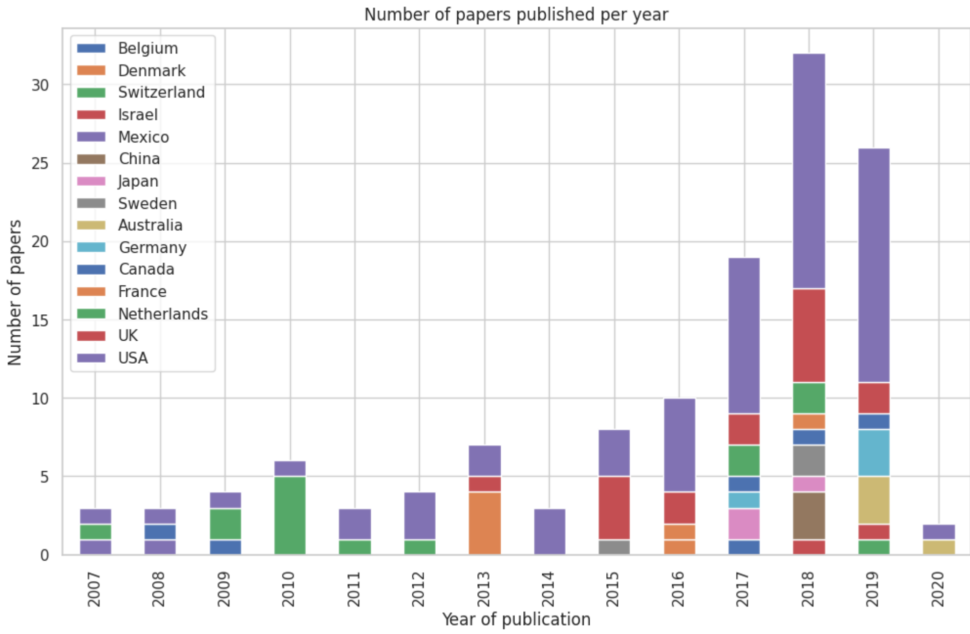


Figure 1: Number of papers published per year

However, from 2016 on, the number of published papers rose sharply, with 19 studies in 2017 et 32 in 2018. These figures show the growing hype about that Intelligent Agent, with breakthrough such as Google DeepMind's AlphaGo in 2015, autonomous cars made available to the public and international initiatives like the European Union's General Data Protection (GDPR) which gives incentives to meet the urgent demands of AI explainability.

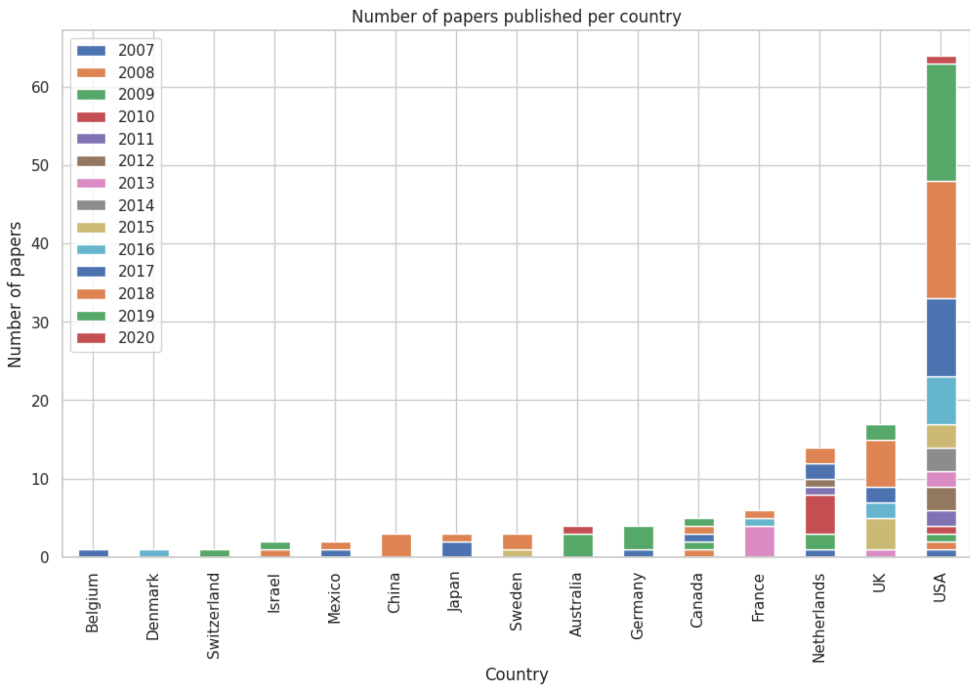


Figure 2: Number of papers published per country

XGDAI research is largely dominated by the United States (63 papers from 2007-2020), followed by the UK (18 studies) and the Netherlands (12 papers). Nonetheless, collaborative studies gathering multiple research institutions have grown in number. It was not shown in the bar chart as the study was attributed to the country with the largest contribution to it.

## RQ 3: Subject

Goal-driven AI systems are either **robots** like those used in factories, deep-sea exploration, home automation and healthcare, or **agents** which do not have a physical body such as recommendation and planning systems.

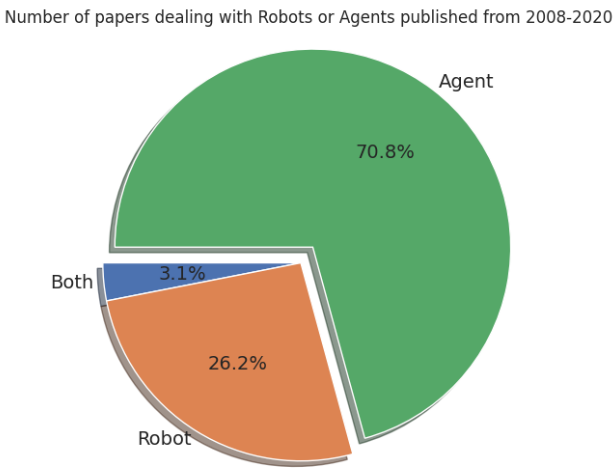


Figure 3: Number of papers dealing with Robots or Agents published from 2008-2020

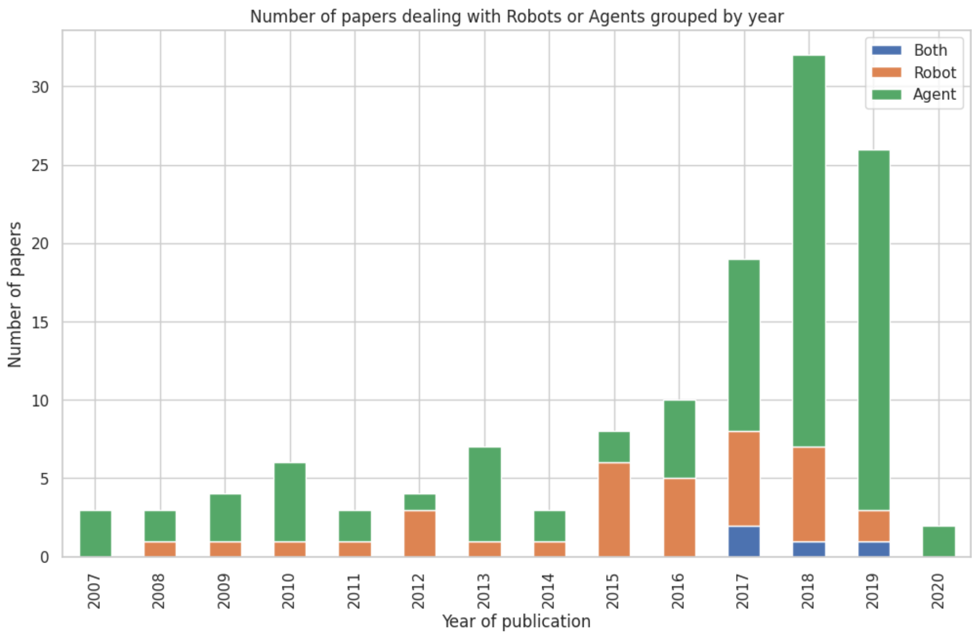


Figure 4: Number of papers dealing with Robots or Agents grouped by year

70.6% of the selected papers concern Agents, 29.4% study Robots. Moreover, the sharp increase in papers from 2017 onwards is actually caused by the augmentation of studies about Agents while the publications about Robots stands at the same level. An explanation could be that Intelligent Agents and most of the studies just focused on how to explain them.

## RQ 4: Recipient

Some of the studies in Explainable Goal-driven AI focus on a agent's/robot's point of view, that is to say generating explanations, understanding the inner workings without really bothering with the explanation communication and reception.

Other study XGDAI from a human's point of view, considering one's feeling and satisfaction towards the justifications. Papers put in this category emphasized evaluations from human users.

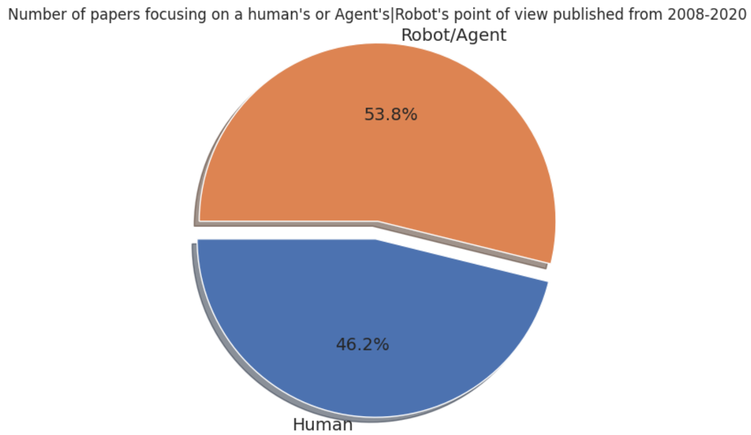


Figure 5: Number of papers focusing on a human's or Agent's/Robot's point of view published from 2008-2020

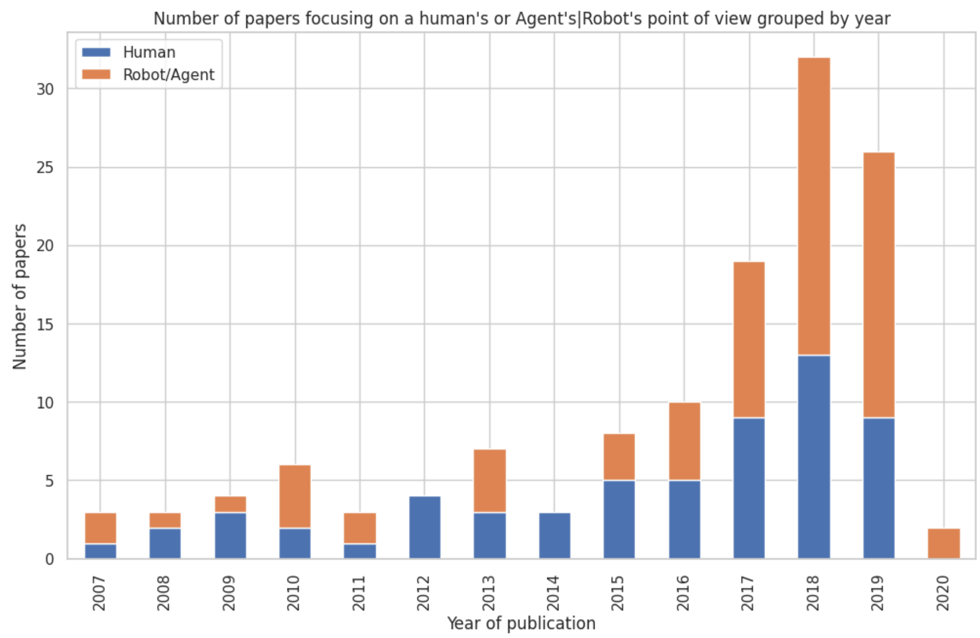


Figure 6: Number of papers focusing on a human's or Agent's/Robot's point of view grouped by year

A majority of the selected papers study XGDAI from a Robot's / Agent's point of view. This shows that the current focus is on Explanation generation rather than Explanation Communication and Explanation reception. It is the first step to produce sound explanations and then make them more easily and efficiently understood.

## RQ 5: Applications scenarii

In this section the different application scenarii mentioned or simulated in the studies are presented. Nonetheless, most of the papers tested their framework in simulations and did not recommend a specific scenario where it can be used.

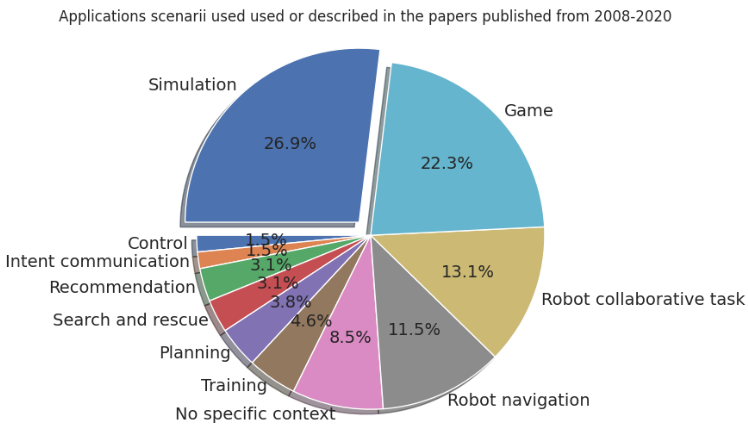


Figure 7: Applications scenarii used used or described in the papers published from 2008-2020

**Simulation (26.9%)**: Situations were simulated in order to assess the performances and the explanations provided by the agent/robot. For example, it could be an everyday life situation with a cooking agent which wants to do pancakes and has to explain the decisions it took at each step [93].

**Game simulation (22.3%)**: More than a fifth of the studies used games as test beds such as Atari games [66] [178] or tried to explain the decisions taken by bots or Non Playable Characters (NPC) [3]. It is a convenient way to assess the performances of the explainable techniques and compare them to black box ones.

**Robot collaborative task (13.1%)**: Studies which focused on robots and humans working together. It could be a situation where a robot passes an object to a human-user in a factory for example [48]. **Robot navigation (11.5%)**: In the concerned papers, a robot is exploring and navigating through obstacles, like an autonomous car would have to [34] [36].

**No specific context (8.5%)**: In these papers, either no application situation was described or the technique can be applied to a lot of applications.

**Training (4.6%)**: Studies in which an agent/robot trains a human, points its mistakes and tells how to correct them. It can be to teach trainees how to fly a plane for example.

**Planning (3.8%)**: In these papers, the agent has to plan the steps in order to achieve a goal, for example optimizing the delivery of parcels.

**Search and rescue (3.1%)**: The agent/robot participates in emergency situations like firefighting or rescue of missing people. **Recommendation (3.1%)**: The agent recommends something to the user, for example why he/she should listen to this song in particular or buy this article.

**Intent communication (1.5%)**: Studies focusing on how to convey the robot's next action or its thinking.

**Control (1.5%)**: Papers studying how explainability can enable a better control or a better debugging of an intelligent agent.

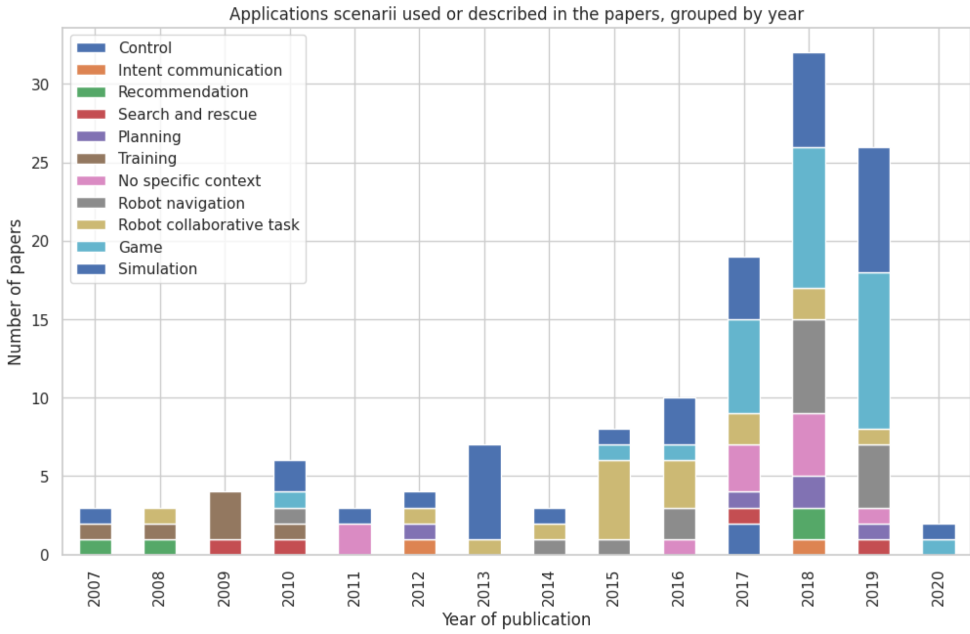


Figure 8: Applications scenarii used or described in the papers, grouped by year

The number of papers using simulations and game simulations as performance test beds are increasing from 2016 onwards. This shows how convenient they are for reproducibility and ease of evaluation.

## RQ 6: Motivations and needs

**Transparency (60.0%)**: Papers focusing on inspecting, understanding and reproducing the inner mechanisms which led to the decisions.

**Trust (32.3%)**: Papers focusing on how to generate coherent explanations and how to present their explanations so that humans would feel confident to trust them.

**Understandability (6.2%)**: Papers focusing on how to generate explanations that can be human-understandable and efficiently understood.

**Intent communication (1.5%)**: Studies focusing on how to convey the robot's next action and/or its thinking.

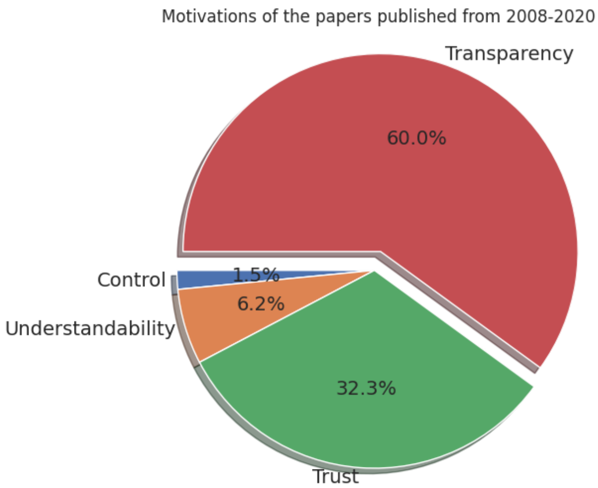


Figure 9: Motivations of the papers published from 2008-2020

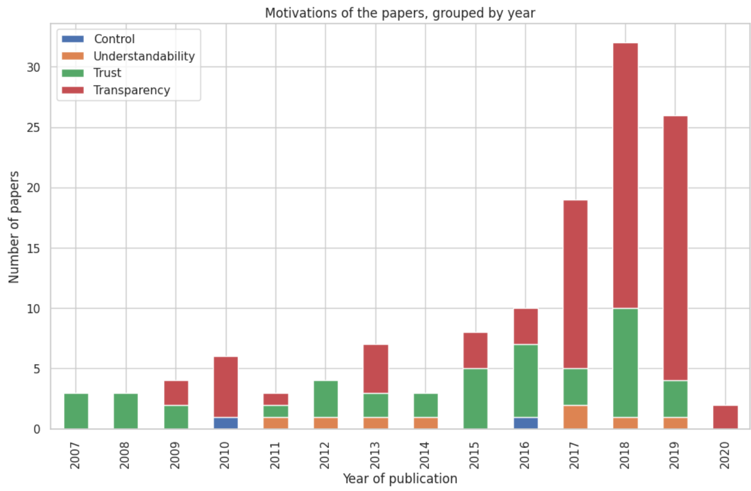


Figure 10: Motivations of the papers, grouped by year

Most of the studies attempt look into the mechanisms which led to a decision (transparency). Figure shows increase of publications about Transparency and not Trust from 2016.

## RQ 7: Social Science and psychological background

**23.8%** of the selected studies clearly mention the **Theory of Mind** when trying to build a framework to efficiently generate and present explanations to a human-user. These ones often describe BDI agents (Belief, Desire, Intention) which can balance the time spent on choosing what do to and executing actions.

**17.7%** of the papers focus on **Folk psychology** to study the reception of the explanation from a human's point of view and conduct empirical evaluations.

However, a majority do not mention either. For example, they can only be interested in explaining the inner model and not really focus in the communication or the reception of the explanations.

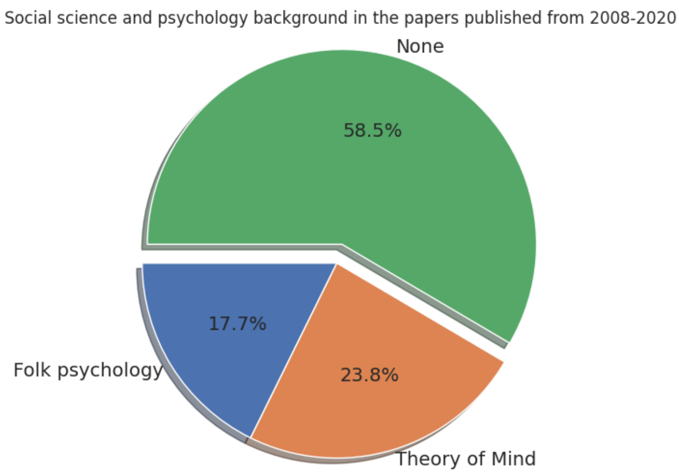


Figure 11: Social science and psychology background in the papers published from 2008-2020

## RQ 8: Types of explanation

There are different kind of explanations that one can expect from an agent or a robot.

**Post-hoc explanations (42.3%)** give justifications without necessarily providing details about the reasoning process that led to the decision.

**Introspective informative explanations (23.8%)** tackle the inner reasoning process of a robot/agent and show what led to a decision.

**Introspective tracing explanations (23.1%)** shed light on the underlying cause of a decision. It is generally used to control the agent/robot's behavior because the observer might be able to trace back a problem or deal with misunderstandings between the system and the user.

**Execution explanations (7.7%)** gives the list of procedures or operations that an agent/robot carried out.

**Contrastive explanations (2.3%)** justify why an event happened or a decision was taken instead of an other one.

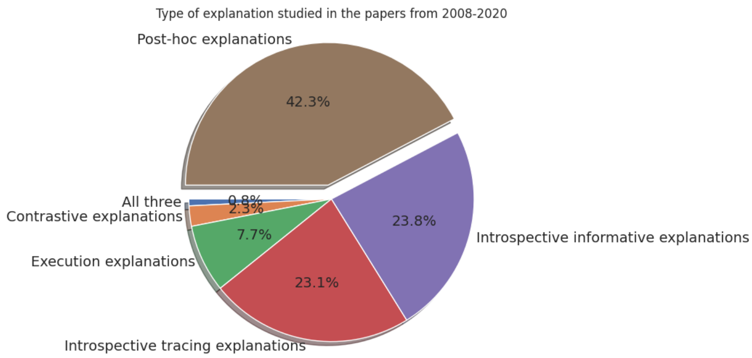


Figure 12: Type of explanation studied in the papers from 2008-2020

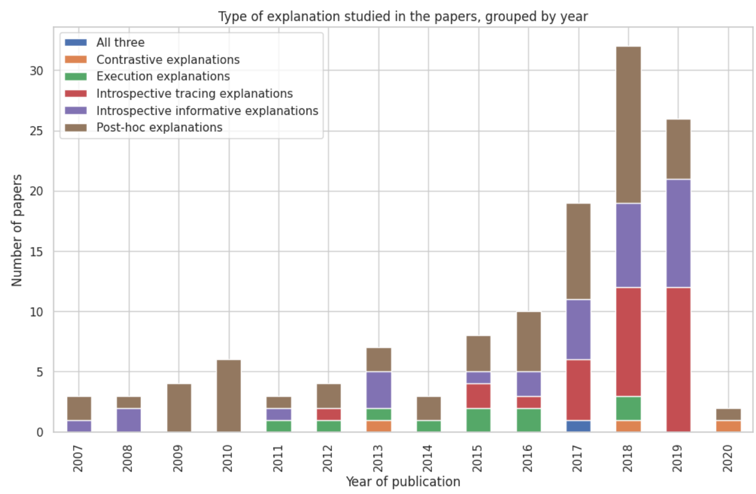


Figure 13: Type of explanation studied in the papers, grouped by year

Most of the papers use post-hoc explanations. This figure can be explained by the fact that papers focusing on human-users generally generate or communicate justifications of actions and not of the inner mechanisms of the AI agent/robot.

## RQ 9: Granularity

One can wonder to what extent is it judicious to explain an intelligent agent/robot. It actually depends on the application of the agent.

**Micro level (40.0%)**: the study explains the inner mechanisms used by the agent. For example, it focuses on the neural network in a model, the most important feature, the reward function or the policy of an RL agent. This level of explanations is judicious for design, control and debugging.

**Macro level (55.4%)**: the study justifies the output decision without inspecting in details the inner thinking of the agent/robot. It could be explaining the planning or the different steps taken by the agent/robot. This level of explanations is judicious for an end-user who does not need technical details of how the system's inner mechanisms but practical justifications, like a worker collaborating with a robot in a factory.

A majority of studies use explanations at a macro level. This figure can be explained by the fact that, in this paper, post-hoc explanations were generally considered as macro level explanations because they usually justify actions and not the inner mechanisms of an agent/robot.

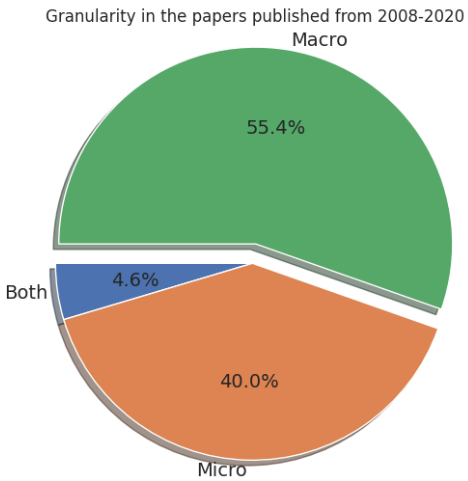


Figure 14: Granularity in the papers published from 2008-2020

## RQ 10: Techniques

The research on Explainable Goal-driven AI is broad in terms of studied techniques, platforms and architectures. Please refer to the mind map to have a synthetic overview. Some techniques are a bit more explored:

- Neural Networks (10.8%)
- Planning (10.0%)
- BDI (10.0%)
- MDP and POMDP (10.0%)
- Causal models (6.2%)
- Saliency maps (3.8%)
- Strategy summarization (3.1%)
- Decision tree (3.1%)

Other techniques include:

- AI rationalization
- Blockchain technology
- Condition random fields
- Contrastive explanations
- Feature visualization
- Framing
- Genetic programming
- Instruction-based behavior
- Interactive RL

- Introspective
- Memory-based RL
- Multi-task RL
- Situation awareness-based Agent Transparency
- Visual Question Answering

Some other studies do not present any technique but focus on an explanation interface (19.2%) that try to display the explanations in a judicious way.

There is not really any trend if we have a look at the techniques presented in the papers, grouped by year. Indeed, XGDAI research is still at its exploratory part and very diverse techniques are studied. However, there is a higher number of papers explaining Neural Network architectures because of their performances and opaque-decision making.

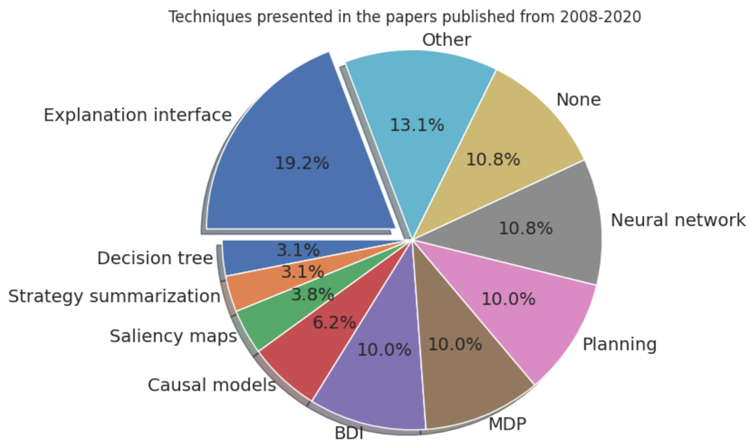


Figure 15: Techniques presented in the papers published from 2008-2020

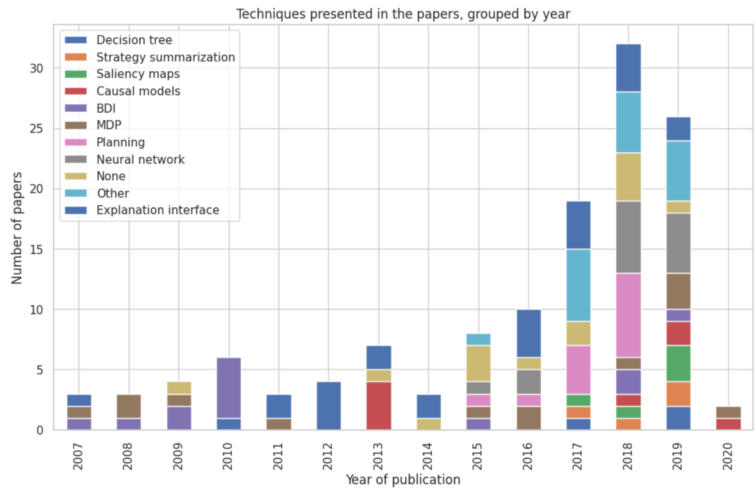


Figure 16: Techniques presented in the papers, grouped by year

## RQ 11: Presentation

This section studies how the explanations are presented to a human-user or a human-observer. It deals with explanation communication.

**Visuals (40.0%)**: Saliency maps, graphs and images.

**Text (31.5%)**: Natural language explanations.

**None**: No specific presentation of the explanations.

**Multi-modal interface (7.7%)**: An interface using different means of communication.

**Expressive motions (6.2%)**: A robot can use motions in order to make it more human-like and show its next action. For example, it can turn its head to a direction to show that it will go this way.

**Logs (2.3%)**: The list of actions or decisions that have been taken.

**Expressive lights (2.3%)**: A robot can use lights to communicate with a human. For instance, it can light the path it will take or project shapes on a screen.

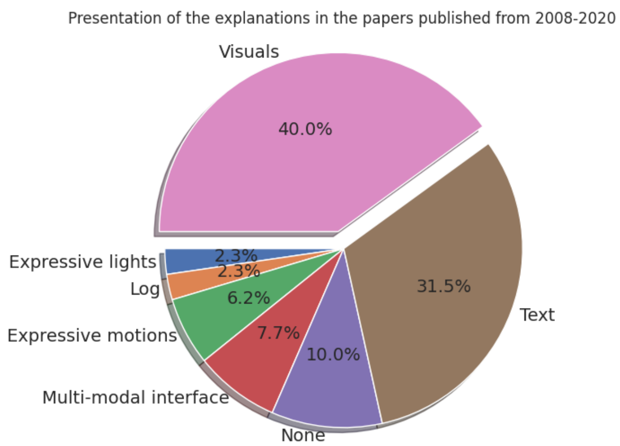


Figure 17: Presentation of the explanations in the papers published from 2008-2020

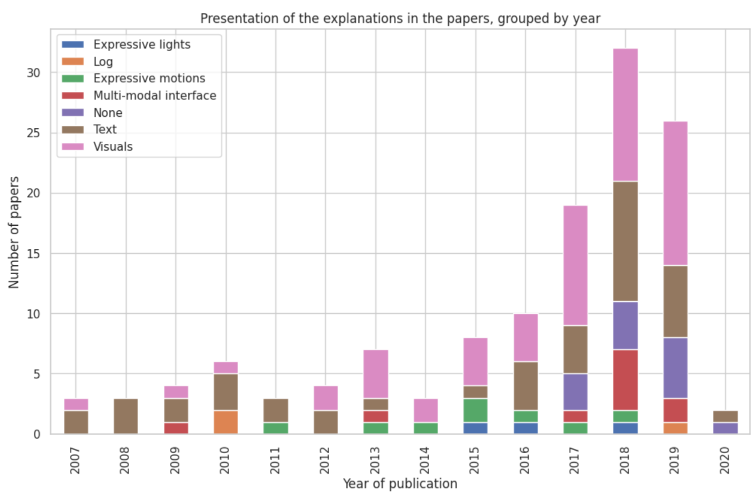


Figure 18: Presentation of the explanations in the papers, grouped by year

Visuals and texts in natural language are the most used channels for presented the explanations because graphs are generally chosen for displaying agent's/robot's performances and text to communicate with a human-user.

## RQ 12: Evaluation

Most of the papers try to assess the validity and the utility of the justifications provided by the agent/robot.

**Test beds (50.8%)**: Evaluate the performances of the agent compared to the usual test beds. The papers usually show that, thanks to their technique, explainability does not trade too much of performances.

**User's understanding (34.6%)**: Studies using polls and empirical studies to evaluate whether a user is satisfied by the explanations and understand them well

**User's feeling (10.8%)**: Studies using polls and empirical studies to evaluate how a human-user perceives the explanations. For example, a robot should not be too intrusive with its means of communication if it is used in a search and rescue mission, for fear of disturbing the mission or simple being annoying.

**Human-likeness (2.3%)**: Studies comparing the generated explanations with how a human would have explained the situation.

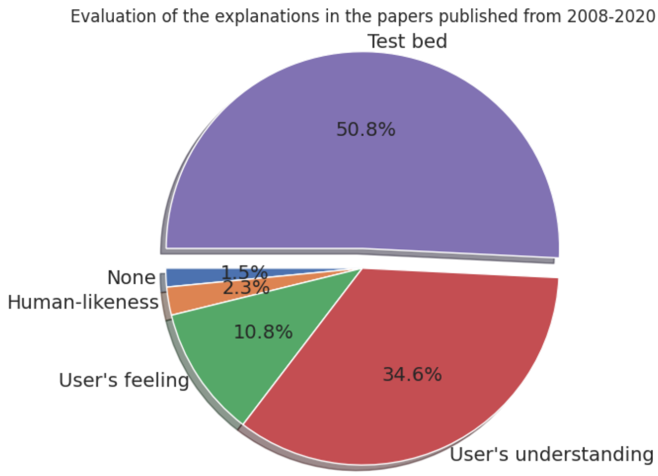


Figure 19: Evaluation of the explanations in the papers, grouped by year

The majority of the studies use tests beds in order to assess the performances of their framework. A large part of them only focus on explanation generation and not explanation communication or personalized explanations yet (user's feeling).

## RQ 13: Future work and challenges

Explainable Goal-driven AI only received attention recently and some studies identify future challenges to improve their work. The papers were divided into categories that were already defined: Explanation generation, Explanation communication, Explanation reception.

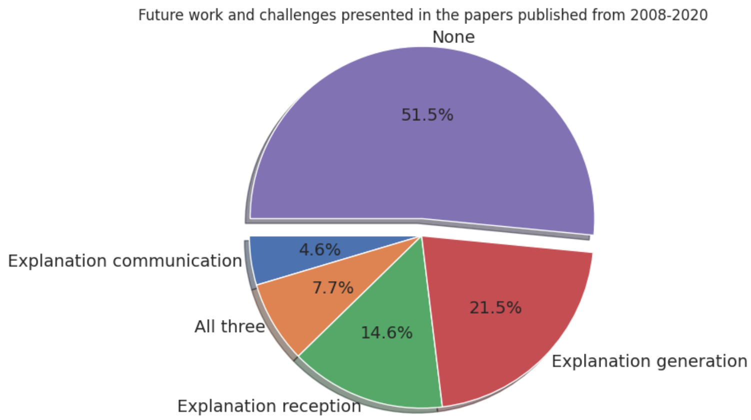


Figure 20: Future work and challenges presented in the papers published from 2008-2020

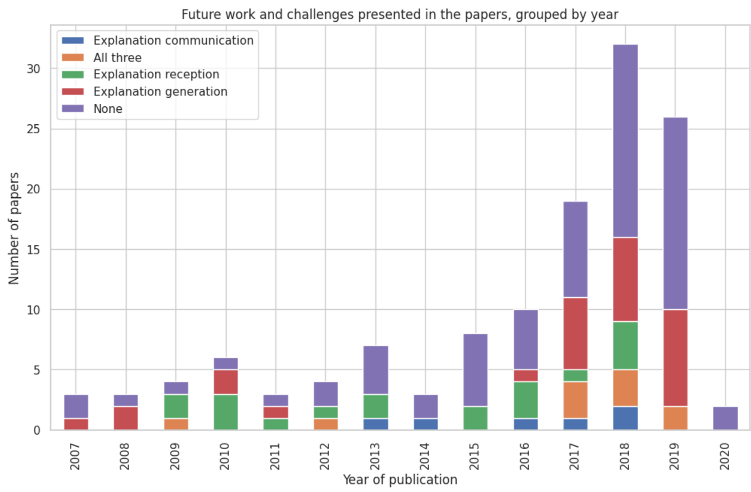


Figure 21: Future work and challenges presented in the papers, grouped by year

Explanation generation was mentioned the most higher number of times as future challenges. This shows that Explainable Goal-driven AI research is still placing more importance on the inner mechanisms of their agent's/robot's model before studying the best way to communicate them. However, they are aware of they lack in the aspect of personalizing of the explanations.

# 3 Road-map for the future challenges

This section gives suggestions that studies can follow in order to tackle the urgent needs for Explainable Goal-driven AI. It follows the three mechanisms previously described.

**Explanation Generation** The decision loops of intelligent agents/robots are becoming more and more complex and more and more efficient in terms of performances. However, most of them do not have any explainability functions. These should be added directly in the loop or should be able to explain the inner workings a posteriori.
There are papers drawing from social and psychological background that study how to formulate a good explanation. However, few concern directly robots and intelligent agents. To generate a dynamic and sound explanation, they should be able to identify the relevant elements for an explanation, figure out its logic and rationales and finally integrate everything

into a sound justification.

**Explanation Communication** Most of the studies detail how an agent/robot presents its explanations to a human-user or a human-observer. However, they often only cover one means of communication (visuals, speech, expressive lights and motions…).
An agent/robot should be able to choose the form of presentation in order to adapt to different situations and different recipients. For this, multi-modal interfaces that can combine channels are needed.

**Explanation Reception** Too few studies are currently evaluating the soundness and relevancy of their agent's/robot's explanations from a user's point of view. Metrics, test beds and methods should be proposed for a clearer and more unified assessment of their techniques.
An agent/robot should have a model that enables it to reflect the user's knowledge evolving and how he/she sees the State of Mind of the intelligent agent

## Conclusion

Driven by the increasing attention of the public and the growing demands concerning ethics, responsibility and the right of explanation, this paper tries to give a global and exhaustive view of the research on Explainable Goal-driven AI in the past ten years. It clarifies the terminology used in the different studies and maps the explanation phases into generation, communication and reception. It then represents the current state of the research by tackling several research questions and providing figures. Eventually, a road-map is proposed in order to better guide the future studies and address the different needs that intelligent agents and robots should be accounted for.

## References

[1] Henny Admoni and Brian Scassellati. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, 6(1):25–63, 2017.

[2] Stephan Alaniz and Zeynep Akata. Xoc: explainable observer-classifier for explainable binary decisions. *arXiv preprint arXiv:1902.01780*, 2019.

[3] Leila Amgoud and Henri Prade. Using arguments for making and explaining decisions. *Artificial Intelligence*, 173(3-4):413–436, 2009.

[4] Dan Amir and Ofra Amir. Highlights: Summarizing agent behavior to people. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1168–1176. International Foundation for Autonomous Agents and Multiagent Systems, 2018.

[5] Ofra Amir, Finale Doshi-Velez, and David Sarne. Agent strategy summarization. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1203–1207. International Foundation for Autonomous Agents and Multiagent Systems, 2018.

[6] Heni Ben Amor, Gerhard Neumann, Sanket Kamthe, Oliver Kroemer, and Jan Peters. Interaction primitives for human-robot cooperation tasks. In *2014 IEEE international conference on robotics and automation (ICRA)*, pages 2831–2837. IEEE, 2014.

[7] Rasmus S Andersen, Ole Madsen, Thomas B Moeslund, and Heni Ben Amor. Projecting robot intentions into human environments. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 294–301. IEEE, 2016.

[8] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. Explainable agents and robots: Results from a systematic literature review. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[9] Raghuram Mandyam Annasamy and Katia Sycara. Towards better interpretability in deep q-networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4561–4569, 2019.

[10] Akanksha Atrey, Kaleigh Clary, and David Jensen. Exploratory not explanatory: Counterfactual analysis of saliency maps for deep reinforcement learning. *arXiv*, pages arXiv–1912, 2019.

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

[12] Joseph Bates et al. The role of emotion in believable agents. *Communications of the ACM*, 37(7):122–125, 1994.

[13] Dianne C Berry, Tony Gillie, and Simon Banbury. What do patients want to know: an empirical approach to explanation generation and validation. *Expert Systems with Applications*, 8(4):419–428, 1995.

[14] Cindy L Bethel. Robots without faces: non-verbal social human-robot interaction. 2009.

[15] Rita Borgo, Michael Cashmore, and Daniele Magazzeni. Towards providing explanations for ai planner decisions. *arXiv preprint arXiv:1810.06338*, 2018.

[16] Nick Bostrom and Eliezer Yudkowsky. The ethics of artificial intelligence. *The Cambridge handbook of artificial intelligence*, 1:316–334, 2014.

[17] Michael W Boyce, Jessie YC Chen, Anthony R Selkowitz, and Shan 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*, pages 179–180, 2015.

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

[19] Daniel Brooks, Abraham Shultz, Munjal Desai, Philip Kovac, and Holly A Yanco. Towards state summarization for autonomous robots. In *2010 AAAI Fall Symposium Series*, 2010.

[20] Davide Calvaresi, Yazan Mualla, Amro Najjar, Stéphane Galland, and Michael Schumacher. Explainable multi-agent systems through blockchain technology. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, pages 41–58. Springer, 2019.

[21] David Cameron, Emily C Collins, Adriel Chua, Samuel Fernando, Owen McAree, Uriel Martinez-Hernandez, Jonathan M Aitken, Luke Boorman, and James Law. Help! i can't reach the buttons: Facilitating helping behaviors towards robots. In *Conference on Biomimetic and Biohybrid Systems*, pages 354–358. Springer, 2015.

[22] Ravi Teja Chadalavada, Henrik Andreasson, Robert Krug, and Achim 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)*, pages 1–6. IEEE, 2015.

[23] Tathagata Chakraborti, Kshitij P Fadnis, Kartik Talamadupula, Mishal Dholakia, Biplav Sri-

vastava, Jeffrey O Kephart, and Rachel KE Bellamy. Visualizations for an explainable planning agent. *arXiv preprint arXiv:1709.04517*, 2017.

[24] Tathagata Chakraborti, Anagha Kulkarni, Sarath Sreedharan, David E Smith, and Subbarao Kambhampati. Explicability? legibility? predictability? transparency? privacy? security? the emerging landscape of interpretable agent behavior. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 29, pages 86–96, 2019.

[25] Tathagata Chakraborti, Sarath Sreedharan, and Subbarao Kambhampati. Explicability versus explanations in human-aware planning. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 2180–2182. International Foundation for Autonomous Agents and Multiagent Systems, 2018.

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

[27] Arjun Chandrasekaran, Deshraj Yadav, Prithvijit Chattopadhyay, Viraj Prabhu, and Devi Parikh. It takes two to tango: Towards theory of ai's mind. *arXiv preprint arXiv:1704.00717*, 2017.

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

[29] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016.

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

[31] Noel CF Codella, Michael Hind, Karthikeyan Natesan Ramamurthy, Murray Campbell, Amit Dhurandhar, Kush R Varshney, Dennis Wei, and Aleksandra Mojsilovic. Teaching meaningful explanations. *arXiv preprint arXiv:1805.11648*, 2018.

[32] European Commission. White paper on artificial intelligence. 2020.

[33] Youri Coppens, Kyriakos Efthymiadis, Tom Lenaerts, Ann Nowé, Tim Miller, Rosina Weber, and Daniele Magazzeni. Distilling deep reinforcement learning policies in soft decision trees. In *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*, pages 1–6, 2019.

[34] Francisco Cruz, Richard Dazeley, and Peter Vamplew. Memory-based explainable reinforcement learning. In *Australasian Joint Conference on Artificial Intelligence*, pages 66–77. Springer, 2019.

[35] Francisco Cruz, Sven Magg, Yukie Nagai, and Stefan Wermter. Improving interactive reinforcement learning: What makes a good teacher? *Connection Science*, 30(3):306–325, 2018.

[36] Alexander G Cunningham, Enric Galceran, Ryan M Eustice, and Edwin Olson. Mpdm: Multipolicy decision-making in dynamic, uncertain environments for autonomous driving. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1670–1677. IEEE, 2015.

[37] Dustin Dannenhauer, Michael W Floyd, Matthew Molineaux, and David W Aha. Learning from exploration: Towards an explainable goal reasoning agent. 2018.

[38] Kerstin Dautenhahn. Socially intelligent robots: dimensions of human–robot interaction. *Philosophical transactions of the royal society B: Biological sciences*, 362(1480):679–704, 2007.

[39] Maartje MA De Graaf and Bertram F Malle. How people explain action (and autonomous intelligent systems should too). In *2017 AAAI Fall Symposium Series*, 2017.

[40] Ewart J de Visser, Marvin Cohen, Amos Freedy, and Raja Parasuraman. A design methodology for trust cue calibration in cognitive agents. In *International conference on virtual, augmented and mixed reality*, pages 251–262. Springer, 2014.

[41] Sandra Devin and Rachid Alami. An implemented theory of mind to improve human-robot shared plans execution. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 319–326. IEEE, 2016.

[42] Thomas Dodson, Nicholas Mattei, and Judy Goldsmith. A natural language argumentation interface for explanation generation in markov decision processes. In *International Conference on Algorithmic DecisionTheory*, pages 42–55. Springer, 2011.

[43] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 0210–0215. IEEE, 2018.

[44] Anca D Dragan, Shira Bauman, Jodi Forlizzi, and Siddhartha S Srinivasa. Effects of robot motion on human-robot collaboration. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 51–58. IEEE, 2015.

[45] Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 301–308. IEEE, 2013.

[46] Upol Ehsan, Brent Harrison, Larry Chan, and Mark 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*, pages 81–87, 2018.

[47] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark 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*, pages 263–274, 2019.

[48] Francisco Elizalde, L Enrique Sucar, Manuel Luque, J Diez, and Alberto Reyes. Policy explanation in factored markov decision processes. In *Proceedings of the 4th European workshop on probabilistic graphical models (PGM 2008)*, pages 97–104, 2008.

[49] Francisco Elizalde, Luis Enrique Sucar, Alberto Reyes, and Pablo Debuen. An mdp approach for explanation generation. In *ExaCt*, pages 28–33, 2007.

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

[51] Yosuke Fukuchi, Masahiko Osawa, Hiroshi Yamakawa, and Michita Imai. Autonomous self-explanation of behavior for interactive reinforcement learning agents. In *Proceedings of the 5th International Conference on Human Agent Interaction*, pages 97–101, 2017.

[52] R Gall. Machine learning explainability vs interpretability: Two concepts that could help restore trust in ai. *KDnuggets News*, 19(1), 2019.

[53] Francisco J Chiyah Garcia, David A Robb, Xingkun Liu, Atanas Laskov, Pedro Patron, and Helen Hastie. Explain yourself: A natural language interface for scrutable autonomous robots. *arXiv preprint arXiv:1803.02088*, 2018.

[54] Marta Garnelo, Kai Arulkumaran, and Murray Shanahan. Towards deep symbolic reinforcement learning. *arXiv preprint arXiv:1609.05518*, 2016.

[55] Fabrizio Ghiringhelli, Jérôme Guzzi, Gianni A Di Caro, Vincenzo Caglioti, Luca M Gambardella, and Alessandro Giusti. Interactive augmented reality for understanding and analyzing multi-robot systems. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1195–1201. IEEE, 2014.

[56] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An approach to evaluating interpretability of machine learning. *arXiv preprint arXiv:1806.00069*, 2018.

[57] Alyssa Glass, Deborah L McGuinness, and Michael Wolverton. Toward establishing trust in adaptive agents. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pages 227–236, 2008.

[58] Ze Gong and Yu Zhang. Behavior explanation as intention signaling in human-robot teaming. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 1005–1011. IEEE, 2018.

[59] Antoine Grea, Laëtitia Matignon, and Samir Aknine. How explainable plans can make planning faster. 2018.

[60] Shirley Gregor and Izak Benbasat. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS quarterly*, pages 497–530, 1999.

[61] Sam Greydanus, Anurag Koul, Jonathan Dodge, and Alan Fern. Visualizing and understanding atari agents. *arXiv preprint arXiv:1711.00138*, 2017.

[62] Victoria Groom and Clifford Nass. Can robots be teammates?: Benchmarks in human–robot teams. *Interaction Studies*, 8(3):483–500, 2007.

[63] William H Guss, Cayden Codel, Katja Hofmann, Brandon Houghton, Noboru Kuno, Stephanie Milani, Sharada Mohanty, Diego Perez Liebana, Ruslan Salakhutdinov, Nicholay Topin, et al. The minerl competition on sample efficient reinforcement learning using human priors. *arXiv preprint arXiv:1904.10079*, 2019.

[64] William H Guss, Brandon Houghton, Nicholay Topin, Phillip Wang, Cayden Codel, Manuela Veloso, and Ruslan Salakhutdinov. Minerl: a large-scale dataset of minecraft demonstrations. *arXiv preprint arXiv:1907.13440*, 2019.

[65] Matthew Guzdial, Joshua Reno, Jonathan Chen, Gillian Smith, and Mark Riedl. Explainable pcgml via game design patterns. *arXiv preprint arXiv:1809.09419*, 2018.

[66] Marc Hanheide, Moritz Göbelbecker, Graham S Horn, Andrzej Pronobis, Kristoffer Sjöö, Alper Aydemir, Patric Jensfelt, Charles Gretton, Richard Dearden, Miroslav Janicek, et al. Robot task planning and explanation in open and uncertain worlds. *Artificial Intelligence*, 247:119–150, 2017.

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

[68] Maaike Harbers, Joost Broekens, Karel Van Den Bosch, and John-Jules Meyer. Guidelines for developing explainable cognitive models. In *Proceedings of ICCM*, pages 85–90. Citeseer, 2010.

[69] Maaike Harbers, John-Jules Meyer, and Karel Van den Bosch. Explaining simulations through self explaining agents. *Journal of Artificial Societies and social simulation*, 12(3):6, 2009.

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

[71] Maaike Harbers, Karel van den Bosch, and John-Jules Meyer. Design and evaluation of explainable bdi agents. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 2, pages 125–132. IEEE, 2010.

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

[73] Maaike Harbers, Karel van den Bosch, and John-Jules Ch Meyer. A theoretical framework for explaining agent behavior. In *SIMULTECH*, pages 228–231, 2011.

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

[75] Helen Hastie, Francisco J Chiyah Garcia, David A Robb, Atanas Laskov, and Pedro Patron. Miriam: A multimodal interface for explaining the reasoning behind actions of remote autonomous systems. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 557–558, 2018.

[76] Helen Hastie, Katrin Lohan, Mike Chantler, David A Robb, Subramanian Ramamoorthy, Ron Petrick, Sethu Vijayakumar, and David Lane. The orca hub: Explainable offshore robotics through intelligent interfaces. *arXiv preprint arXiv:1803.02100*, 2018.

[77] Bradley Hayes and Julie A Shah. Improving robot controller transparency through autonomous policy explanation. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, pages 303–312. IEEE, 2017.

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

[79] Aroua Hedhili, Wided Lejouad Chaari, and Khaled Ghédira. Causal maps for explanation in multi-agent system. In *Intelligent Informatics*, pages 183–191. Springer, 2013.

[80] Aroua Hedhili, Wided Lejouad Chaari, and Khaled Ghédira. Explanation language syntax for multi-agent systems. In *2013 World Congress on Computer and Information Technology (WCCIT)*, pages 1–6. IEEE, 2013.

[81] Daniel Hein, Steffen Udluft, and Thomas A Runkler. Interpretable policies for reinforcement learning by genetic programming. *Engineering Applications of Artificial Intelligence*, 76:158–169, 2018.

[82] Thomas Hellström and Suna Bensch. Understandable robots-what, why, and how. *Paladyn, Journal of Behavioral Robotics*, 9(1):110–123, 2018.

[83] Daniel Holliday, Stephanie Wilson, and Simone Stumpf. The effect of explanations on per-

ceived control and behaviors in intelligent systems. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 181–186. 2013.

[84] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, pages 9734–9745, 2019.

[85] Tobias Huber, Dominik Schiller, and Elisabeth André. Enhancing explainability of deep reinforcement learning through selective layer-wise relevance propagation. In *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)*, pages 188–202. Springer, 2019.

[86] Rahul Iyer, Yuezhang Li, Huao Li, Michael Lewis, Ramitha Sundar, and Katia Sycara. Transparency and explanation in deep reinforcement learning neural networks. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 144–150, 2018.

[87] W Lewis Johnson. Agents that explain their own actions. *AD-A280 063*, 8:21, 1994.

[88] W Lewis Johnson. Agents that learn to explain themselves. In *AAAI*, pages 1257–1263, 1994.

[89] Zoe Juozapaitis, Anurag Koul, Alan Fern, Martin Erwig, and Finale Doshi-Velez. Explainable reinforcement learning via reward decomposition. In *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*, pages 47–53, 2019.

[90] Frank Kaptein, Joost Broekens, Koen Hindriks, and Mark 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)*, pages 676–682. IEEE, 2017.

[91] Frank Kaptein, Joost Broekens, Koen Hindriks, and Mark 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)*, pages 88–93. IEEE, 2017.

[92] Saurabh Kaushik. Holy grail of ai for enterprise — explainable ai. 2017.

[93] Omar Zia Khan, Pascal Poupart, and James P Black. Explaining recommendations generated by mdps. In *ExaCt*, pages 13–24, 2008.

[94] Omar Zia Khan, Pascal Poupart, and James P Black. Minimal sufficient explanations for factored markov decision processes. In *Nineteenth International Conference on Automated Planning and Scheduling*, 2009.

[95] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 267–280. Springer, 2019.

[96] Matthew Klenk, Matt Molineaux, and David W Aha. Goal-driven autonomy for responding to unexpected events in strategy simulations. *Computational Intelligence*, 29(2):187–206, 2013.

[97] W Bradley Knox and Peter Stone. Interactively shaping agents via human reinforcement: The tamer framework. In *Proceedings of the fifth international conference on Knowledge capture*, pages 9–16, 2009.

[98] Jeamin Koo, Jungsuk Kwac, Wendy Ju, Martin Steinert, Larry Leifer, and Clifford Nass. Why did my car just do that? explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. *International Journal on Interactive Design and Manufac-*

*turing (IJIDeM)*, 9(4):269–275, 2015.

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

[100] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen 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*, pages 3–10. IEEE, 2013.

[101] Isaac Lage, Daphna Lifschitz, Finale Doshi-Velez, and Ofra Amir. Toward robust policy summarization. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 2081–2083. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[102] Pat Langley, Ben Meadows, Mohan Sridharan, and Dongkyu Choi. Explainable agency for intelligent autonomous systems. In *Twenty-Ninth IAAI Conference*, 2017.

[103] Stefan Larsson and Fredrik Heintz. Transparency in artificial intelligence. *Internet Policy Review*, 9(2), 2020.

[104] Jung Hoon Lee. Complementary reinforcement learning towards explainable agents. *arXiv preprint arXiv:1901.00188*, 2019.

[105] Iolanda Leite, Carlos Martinho, and Ana Paiva. Social robots for long-term interaction: a survey. *International Journal of Social Robotics*, 5(2):291–308, 2013.

[106] Brian Y Lim and Anind 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*, pages 157–166, 2011.

[107] Brian Y Lim, Anind K Dey, and Daniel Avrahami. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2119–2128, 2009.

[108] Zachary Chase Lipton. The myth of interpretability. 2016.

[109] Guiliang Liu, Oliver Schulte, Wang Zhu, and Qingcan Li. Toward interpretable deep reinforcement learning with linear model u-trees. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 414–429. Springer, 2018.

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

[111] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. Explainable reinforcement learning through a causal lens. *arXiv preprint arXiv:1905.10958*, 2019.

[112] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. Distal explanations for explainable reinforcement learning agents. *arXiv preprint arXiv:2001.10284*, 2020.

[113] Matthew Mayo. Interpreting machine learning models: An overview. 2017.

[114] Sean McGregor, Hailey Buckingham, Thomas G Dietterich, Rachel Houtman, Claire Montgomery, and Ronald Metoyer. Facilitating testing and debugging of markov decision processes with interactive visualization. In *2015 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 53–61. IEEE, 2015.

[115] Deborah L McGuinness, Alyssa Glass, Michael Wolverton, and Paulo Pinheiro Da Silva. A categorization of explanation questions for task processing systems. In *ExaCt*, pages 42–48,

2007.

[116] Deborah L McGuinness, Alyssa Glass, Michael Wolverton, and Paulo Pinheiro Da Silva. Explaining task processing in cognitive assistants that learn. In *AAAI spring symposium: Interaction challenges for intelligent assistants*, pages 80–87, 2007.

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

[118] Stephanie Milani, Nicholay Topin, Brandon Houghton, William H Guss, Sharada P Mohanty, Oriol Vinyals, and Noboru Sean Kuno. The minerl competition on sample-efficient reinforcement learning using human priors: A retrospective. *arXiv preprint arXiv:2003.05012*, 2020.

[119] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[120] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[121] Matthew Molineaux, Dustin Dannenhauer, and David W Aha. Towards explainable npcs: a relational exploration learning agent. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[122] Alexander Mott, Daniel Zoran, Mike Chrzanowski, Daan Wierstra, and Danilo Jimenez Rezende. Towards interpretable reinforcement learning using attention augmented agents. In *Advances in Neural Information Processing Systems*, pages 12329–12338, 2019.

[123] Bonnie M Muir. Trust between humans and machines, and the design of decision aids. *International journal of man-machine studies*, 27(5-6):527–539, 1987.

[124] Clifford Nass, Ing-Marie Jonsson, Helen Harris, Ben Reaves, Jack Endo, Scott Brave, and Leila Takayama. Improving automotive safety by pairing driver emotion and car voice emotion. In *CHI'05 extended abstracts on Human factors in computing systems*, pages 1973–1976, 2005.

[125] Mark A Neerincx, Jasper van der Waa, Frank Kaptein, and Jurriaan van Diggelen. Using perceptual and cognitive explanations for enhanced human-agent team performance. In *International Conference on Engineering Psychology and Cognitive Ergonomics*, pages 204–214. Springer, 2018.

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

[127] Mayada Oudah, Talal Rahwan, Tawna Crandall, and Jacob W Crandall. How ai wins friends and influences people in repeated games with cheap talk. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[128] Prajwal Paudyal. Should ai explain itself? or should we design explainable ai so that it doesn't have to. 2019.

[129] Sean Penney, Jonathan Dodge, Claudia Hilderbrand, Andrew Anderson, Logan Simpson, and Margaret Burnett. Toward foraging for understanding of starcraft agents: An empirical study. In *23rd International Conference on Intelligent User Interfaces*, pages 225–237, 2018.

[130] Joëlle Pineau. Mooc university of alberta, fundamentals of reinforcement learning. 2018.

[131] Rey Pocius, Lawrence Neal, and Alan Fern. Strategic tasks for explainable reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 10007–10008, 2019.

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

[133] Stephanie Rosenthal, Sai P Selvaraj, and Manuela M Veloso. Verbalization: Narration of autonomous robot experience. In *IJCAI*, pages 862–868, 2016.

[134] Aaron M Roth, Nicholay Topin, Pooyan Jamshidi, and Manuela Veloso. Conservative q-improvement: Reinforcement learning for an interpretable decision-tree policy. *arXiv preprint arXiv:1907.01180*, 2019.

[135] Christian Rupprecht, Cyril Ibrahim, and Chris Pal. Visualizing and discovering behavioural weaknesses in deep reinforcement learning. 2018.

[136] Christian Rupprecht, Cyril Ibrahim, and Christopher J Pal. Finding and visualizing weaknesses of deep reinforcement learning agents. *arXiv preprint arXiv:1904.01318*, 2019.

[137] Fatai Sado, Chu Kiong Loo, Matthias Kerzel, and Stefan Wermter. Explainable goal-driven agents and robots–a comprehensive review and new framework. *arXiv preprint arXiv:2004.09705*, 2020.

[138] Wojciech Samek. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature, 2019.

[139] Vieri Giuliano Santucci, Gianluca Baldassarre, and Marco Mirolli. Grail: a goal-discovering robotic architecture for intrinsically-motivated learning. *IEEE Transactions on Cognitive and Developmental Systems*, 8(3):214–231, 2016.

[140] Aroua Hedhili Sbaï and Wided Lejouad Chaari. Extended causal map for reasoning explanation in multi–agent systems. *International Journal of Intelligent Systems Technologies and Applications*, 12(3-4):301–315, 2013.

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

[142] Bastian Seegebarth, Felix Müller, Bernd Schattenberg, and Susanne 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.

[143] Pedro Sequeira, Eric Yeh, and Melinda T Gervasio. Interestingness elements for explainable reinforcement learning through introspection. In *IUI Workshops*, page 7, 2019.

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

[145] Tianmin Shu, Caiming Xiong, and Richard Socher. Hierarchical and interpretable skill acquisition in multi-task reinforcement learning. *arXiv preprint arXiv:1712.07294*, 2017.

[146] Shirin Sohrabi, Jorge A Baier, and Sheila A McIlraith. Preferred explanations: Theory and generation via planning. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

[147] Sichao Song and Seiji 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*, pages 1–6, 2018.

[148] Sarath Sreedharan, Subbarao Kambhampati, et al. Balancing explicability and explanation in human-aware planning. In *2017 AAAI Fall Symposium Series*, 2017.

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

[150] Xiaomeng Su, Mihhail Matskin, and Jinghai Rao. Implementing explanation ontology for agent system. In *Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003)*, pages 330–336. IEEE, 2003.

[151] Roykrong Sukkerd, Reid Simmons, and David Garlan. Toward explainable multi-objective probabilistic planning. In *2018 IEEE/ACM 4th International Workshop on Software Engineering for Smart Cyber-Physical Systems (SEsCPS)*, pages 19–25. IEEE, 2018.

[152] Roykrong Sukkerd, Reid Simmons, and David Garlan. Tradeoff-focused contrastive explanation for mdp planning. *arXiv preprint arXiv:2004.12960*, 2020.

[153] William Swartout, Cecile Paris, and Johanna Moore. Explanations in knowledge systems: Design for explainable expert systems. *IEEE Expert*, 6(3):58–64, 1991.

[154] Aaquib Tabrez and Bradley Hayes. Improving human-robot interaction through explainable reinforcement learning. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 751–753. IEEE, 2019.

[155] Leila Takayama, Doug Dooley, and Wendy Ju. Expressing thought: improving robot readability with animation principles. In *Proceedings of the 6th international conference on Human-robot interaction*, pages 69–76, 2011.

[156] Andreas Theodorou, Robert H Wortham, and Joanna J Bryson. Why is my robot behaving like that? designing transparency for real time inspection of autonomous robots. In *AISB Workshop on Principles of Robotics*. University of Bath, 2016.

[157] Nava Tintarev and Judith Masthoff. A survey of explanations in recommender systems. In *2007 IEEE 23rd international conference on data engineering workshop*, pages 801–810. IEEE, 2007.

[158] Nicholay Topin and Manuela Veloso. Generation of policy-level explanations for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2514–2521, 2019.

[159] Jasper van der Waa, Jurriaan van Diggelen, Karel van den Bosch, and Mark Neerincx. Contrastive explanations for reinforcement learning in terms of expected consequences. *arXiv preprint arXiv:1807.08706*, 2018.

[160] Abhinav Verma, Vijayaraghavan Murali, Rishabh Singh, Pushmeet Kohli, and Swarat Chaudhuri. Programmatically interpretable reinforcement learning. *arXiv preprint arXiv:1804.02477*, 2018.

[161] Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani, Heinrich Küttler, John Agapiou, Julian Schrittwieser, et al. Starcraft ii: A new challenge for reinforcement learning. *arXiv preprint arXiv:1708.04782*, 2017.

[162] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. Transparent, explainable, and accountable ai for robotics. 2017.

[163] Ning Wang, David V Pynadath, and Susan 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*, pages 997–1005, 2016.

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

[165] Xiting Wang, Yiru Chen, Jie Yang, Le Wu, Zhengtao Wu, and Xing Xie. A reinforcement learning framework for explainable recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 587–596. IEEE, 2018.

[166] Handy Wicaksono and Claude Sammut1 Raymond Sheh. Towards explainable tool creation by a robot. In *IJCAI-17 Workshop on Explainable AI (XAI)*, page 63, 2017.

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

[168] Robert H Wortham and Andreas Theodorou. Robot transparency, trust and utility. *Connection Science*, 29(3):242–248, 2017.

[169] Robert H Wortham, Andreas Theodorou, and Joanna 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.

[170] Robert H Wortham, Andreas Theodorou, and Joanna 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)*, pages 1424–1431. IEEE, 2017.

[171] Rosemarie E Yagoda and Douglas J Gillan. You want me to trust a robot? the development of a human–robot interaction trust scale. *International Journal of Social Robotics*, 4(3):235–248, 2012.

[172] Yu Zhang, Sarath Sreedharan, Anagha Kulkarni, Tathagata Chakraborti, Hankz Hankui Zhuo, and Subbarao Kambhampati. Plan explicability and predictability for robot task planning. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 1313–1320. IEEE, 2017.

[173] Guoshuai Zhao, Hao Fu, Ruihua Song, Tetsuya Sakai, Xing Xie, and Xueming Qian. Why you should listen to this song: Reason generation for explainable recommendation. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1316–1322. IEEE, 2018.