

# Trust in Human Machine Partnerships

Gerard Canal  
Dong Huynh  
Senka Krivic  
Quratul-ain Mahesar  
Department of Informatics, King's  
College London

Rita Borgo  
Andrew Coles  
Luc Moreau  
Department of Informatics, King's  
College London

Menisha Patel  
Paul Luff  
King's Business School, King's  
College London

Archie Drake  
Perry Keller  
The Dickson Poon School of Law,  
King's College London

Simon Parsons  
School of Computer Science,  
University of Lincoln

## ABSTRACT

We are working on components that can compute explanations for a model-based AI, in particular a system for AI planning. This short paper sketches the components of the system and discusses some issues that arise when considering how to compute explanations.

### ACM Reference Format:

Gerard Canal, Dong Huynh, Senka Krivic, Quratul-ain Mahesar, Rita Borgo, Andrew Coles, Luc Moreau, Menisha Patel, Paul Luff, Archie Drake, Perry Keller, and Simon Parsons. 2020. Trust in Human Machine Partnerships. In *Proceedings of WebSci'20 Workshop: Explanations for AI: Computable or Not? (exAI2020)*. ACM, New York, NY, USA, 3 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

## 1 INTRODUCTION

As artificial intelligence (AI) becomes increasingly widely deployed, the need for AI to support interaction with humans becomes ever more acute. If AI systems are going to work effectively together with humans, then those systems must develop trust in the human-machine team in order that humans are comfortable sharing responsibility with them. We are working from the position that such trust is best established when the humans involved are confident that the AI reaches sensible conclusions because they understand the reasoning involved. That is, they understand that i) decisions are based on appropriate information, and that information has been processed in suitable ways; ii) that the reasons behind the decisions are communicated clearly and effectively; and iii) they can engage in a process of discussing and questioning decisions with the AI. Together, these components provide a sound basis for a system of *explainable AI* (xAI).

Figure 1 sketches the elements of our work. Given the centrality of AI planning, our work is firmly positioned in the area of model-based AI. Thus the inputs include domain knowledge as well as

data. The reasoning is performed on this combination. Methods from AI planning use knowledge and data to generate plans, for example for the operation of a robot. Methods from computational *argumentation* use knowledge and data to construct reasons behind the choices made in the plans. Methods from *provenance* track the operations performed on knowledge and data to allow them to be recorded and examined later. The resulting plans, reasons, and provenance traces provide the basis for a range of *explanations* and *visualisations* which a user can navigate. Users are not just passive receivers of explanations, but can also interrogate explanations and navigate visualizations in order to build understanding.

Developing convincing explanations is not just a matter of applying suitable technology. Identifying suitable forms of explanation and justification requires us to consider the social context in which explanation is required, and the very question of decisions being made by autonomous software systems raises many legal issues.

In the rest of this paper, we describe how we are addressing these problems, provide some detail on the technologies that we are using, and discuss how our progress to date highlights issues for the endeavour of computing explanations for AI systems.

## 2 BACKGROUND

In this section, we briefly discuss the various technologies that we are bringing together in this work, and sketch how they are related in our work.

### 2.1 AI Planning

The definition of a decision-making problem as a planning problem introduces a model and structure from which explanations can be generated. Planning tries to identify sets of actions that a given system, such as a mobile robot, can perform to achieve specific goals. More formally, a planning problem is a tuple  $\Pi = \langle S, A, s_0, g \rangle$ , where  $S = P \cup V$  is a set of states made from a set  $P$  of propositions and a vector of real variables (fluents)  $V$ .  $A$  is a set of actions that modify the state;  $s_0 \in S$  describes the *initial state* of the problem; and  $g \in S$  describes the *goal state*. A solution to the problem is a plan detailing *what* actions to do, and *when* to do them, in order to transform the initial state  $s_0$  into the goal state  $g$ . Planning problems span a wide range of areas, from optimisation to robotics, and may involve conflicting objectives and timing constraints.

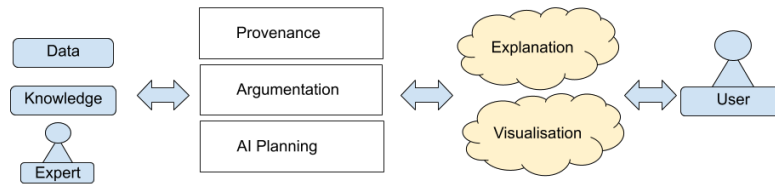
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

exAI2020, Southampton, July 7–8, 2020

© 2020 Association for Computing Machinery.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>



**Figure 1: An overview of the elements we are developing**

The user will have their own assumptions and expectations of what can be done in terms of the actions available to the system for which the plan is being created. Making decisions based on a model will allow showing the reasons behind each of the decisions, as opposed to a black box or data-based manner. Therefore, obtaining and presenting these reasons can be used to build trust, allowing the system to show which are the available options that it has and confront them with user’s ideas.

## 2.2 Provenance

Provenance, defined as a record describing how organisations, people, data, and activities may have influenced a decision or an outcome of some AI [8], can be applied to planning systems. In the planning context, provenance can be employed to track which organisation, people, sensors, execution of a previous plan may have influenced the domain knowledge and the world’s states used in the planning process and, eventually, the resulting plan and its likelihood to succeed. For instance, an engineer may have updated the domain knowledge with a new constraint; a delayed execution of a previous plan may lead to a world’s state that is different to the expectation of a stakeholder; or, fuel/battery updates from robots may dictate which are viable assets to be deployed. During the execution of a plan, situations may, and often do, change: a robot may encounter an obstruction in the real world that was not aware by the planning system, making it impossible to achieve its goal; an operator may also add new goals or override the urgency of existing ones, affecting how/whether the plan will be completed. Therefore, tracking all the inputs, human or otherwise, fed into the planning process, how they are aggregated or transformed, and pertinent facts in a dynamic executing environment is crucial in order to establish dependencies and responsibilities to help with explaining planning decisions.

Recording the provenance of a plan and plan execution provides us with an audit trail to enable tracing back all the influences that went into the generation of the plan and what may have impacted its execution. Based on the provenance, the robustness of the plan can be verified (manually or automatically) and explained to its stakeholders: which inputs it depends on, when they were updated, how they were processed, and who/what is responsible for which input. During plan execution, changes in the inputs that may affect a plan can be monitored to determine if it is still viable and whether re-planning is required [9]. When a plan is aborted or unduly delayed, an explanation could be derived from a comparison of the current world’s states and those recorded in the provenance. In summary, the provenance of planning processes is a basis to help us investigate how they took place and, combining with work in explainable

planning [4], better understand decisions made by human-machine planning systems.

## 2.3 Argumentation

Argumentation [10] is a logical model of reasoning that has its origins in philosophy. Work on computational argumentation, first started appearing in the second half of the 1980s, and argumentation is now well established as an important sub-field within artificial intelligence. It provides a mechanism for the evaluation of possible conclusions or claims by considering reasons for and against them. These reasons, i.e., arguments and counter-arguments, provide support for and against the conclusions or claims, through a combination of dialectical and logical reasoning.

Argumentation is connected to the idea of establishing trust in AI systems by explaining the results and processes of the computation of a solution or decision. In this work, we are applying the same process to planning. So far we have developed two mechanisms by which argumentation can do this. First, we have developed a mechanism for taking the plan output by a planner and constructing reasons for every step (action) in the plan. These reasons are given by viewing a plan step as being a transformation from one state to another, generating a functional explanation in the same style as [6]. Secondly, we have introduced argumentation schemes [1] that can be used generate fuller explanations in this domain. The schemes allow the arguments (reasons) for plan steps to be expressed in natural language, and each scheme (which explains one aspect of a plan step) is associated with a set of critical questions which enumerate the reasons why steps might not be appropriate. These questions can be used to prompt user reflection on the suitability of a particular plan.

## 2.4 Explanations

We aim to facilitate the notion of users interacting with the planning process, as suggested by [11], allowing them to seek additional information about the planner and explanations for its decisions. In this way, users see themselves as collaborators with the planner.

Plan explanations must address questions from several perspectives of potential users of the system. There will be experts questioning the system as well as lay users. Therefore we are providing explanations in different ways: contrastive explanations [4] – providing users counterfactual examples of plans considering users questions; planning justifications – providing reasons for changes in the planning process; ethical comparison of plans [7]; argumentation based explanations – providing arguments expressed in natural language [1]; provenance explanations – using provenance

to examine where information used for planning came from and how it affected the planning process.

## 2.5 Visualisation

Visualization plays a pivotal role in supporting explainability, a core element of Trust, in the context of AI as either a mean to provide a communication/interaction channel between AI agent(s) and end user(s) or as an exploration tool to dive inside its decision flow. Several attempts have been made to develop platforms to support the latter [5] but progress is still needed on the former. In [2] a first attempt is made to enhance the traditional interface based communication approach to enrich the conversation flow and embed the user within the decision making process. The context of human-machine partnership however is more complex as it deals with both continuous and discrete events prompting the need for adaptable layouts capable of leveraging both context, objective as in final goal and tasks needed to achieve such goal, as well as human abilities and know-how. A core starting point is looking at the argumentation flow within the different level of conversation that model interaction and exchange in a partnership. Based on works from [3] we focus our attention on the discrete nature of these phenomena, the level of abstraction needed to express the different layers of arguments inner-workings, and the semantics of an argumentation flow. Starting from arguments parameter space it is possible to move towards the construction of parameterized visual abstractions to express the argument space and visual summaries of its specialization. Core to the creation of visual abstraction is the principle of visual analytics as human in the loop visualization which leverages and favours human perceptual and cognitive capabilities as well as level of expertise.

## 3 SOCIAL AND ORGANIZATIONAL ASPECTS

To develop trustworthy systems we need to consider the social and organisational context in which such systems will be embedded. In this regard we are undertaking two kinds of studies. The first focusses on the experts and those that inform the content of the model. The second considers the user and how they might interact with the outcomes of the system. To provide concrete motivating examples we are collaborating with organisations with very different types of context: one already uses AI and planning systems for a very specific task and the other has a number of requirements for which AI systems maybe appropriate. By considering these two different contexts we can explore the barriers and opportunities of integrating argumentation, provenance and planning. Hence, we investigate the kinds of explanations that are currently being used, the practical resources which are used to produce them and how people characterise what constitutes a “good” explanation. We are particularly interested in how the organisational context both shapes the requirements for explanations and also how they are produced.

Our approach draws from qualitative social scientific approaches. For example, we undertake detailed semi-structured interviews with experts and potential users to elicit their experience and perceptions regarding planning and explanations. These focus on the nature of explanations in organisational contexts and the practical constraints in which they are produced and understood. We also

plan to undertake distinctive studies for assessing the techniques and models we develop. These will include quasi-naturalistic studies of prototypes, qualitative experiments where we will investigate breaches of trust and public engagement activities where we demonstrate our approaches to a broader audience. In this way our team is developing a distinctive multi-disciplinary approach. There are considerable challenges in undertaking this form of collaboration.

## 4 SUMMARY

Our work so far suggests that there are several forms of explanation for AI planning that are computable. As in [4], we can produce contrastive explanations, and we can also generate explanations that take a more causal direction [1]. While such explanations appeal to some users, it is not clear that they appeal to all users. In other words, while we can generate such explanations, they may not be useful. Identifying whether or not they are useful is ongoing work. The effectiveness of the explanations that we can currently compute will be tested through user studies, developed around a scenario involving a mobile robot operating in an office environment. In parallel, the work laid out in Section 3 aims to help us understand what kinds of explanation will help the groups of users who are engaged with our project. If it turns out that the contrastive and causal explanations that we can currently compute are not sufficient, then we will look to construct computable explanations that are more useful. Finally, we need to understand better what the legal requirements of explanations are, and how these might be satisfied. This is also ongoing work.

## ACKNOWLEDGMENTS

This work was partially supported by EPSRC grant EP/R033722/1.

## REFERENCES

- [1] Quratul ain Mahesar and Simon Parsons. 2020. Argument Schemes for Explainable Planning. *arXiv:cs.AI/2005.05849*
- [2] Rita Borgo, Michael Cashmore, and Daniele Magazzeni. 2018. Towards Providing Explanations for AI Planner Decisions. *CoRR* abs/1810.06338 (2018).
- [3] Rita Borgo, Johannes Kehr, David Chung, Eamonn Maguire, Robert Laramée, Helwig Hauser, Matthew Ward, and Min Chen. 2013. Glyph-based Visualization: Foundations, Design Guidelines, Techniques and Applications.
- [4] Michael Cashmore, Anna Collins, Benjamin Krarup, Senka Krivic, Daniele Magazzeni, and David Smith. 2019. Towards Explainable Planning as a Service. In *ICAPS-19 Workshop on Explainable Planning*.
- [5] Angelos Chatzimpampas, Rafael M. Martins, Ilir Jusufi, and Andreas Kerren. 0. A survey of surveys on the use of visualization for interpreting machine learning models. *Information Visualization* 0, 0 (0), 1473871620904671.
- [6] Xiuyi Fan. 2018. On generating explainable plans with assumption-based argumentation. In *International Conference on Principles and Practice of Multi-Agent Systems*. Springer, 344–361.
- [7] Benjamin Krarup, Senka Krivic, Felix Lindner, and Derek Long. 2020. Towards Contrastive Explanations for Comparing the Ethics of Plans. In *ICRA-20 Workshop on Against Robot Dystopias*.
- [8] Luc Moreau and Paolo Missier. 2013. *PROV-DM: The PROV Data Model*. Technical Report. World Wide Web Consortium. <http://www.w3.org/TR/2013/REC-prov-dm-20130430/> W3C Recommendation.
- [9] Sarvapali D. Ramchurn, Trung Dong Huynh, Feng Wu, Yukki Ikuno, Jack Flann, Luc Moreau, Joel E. Fischer, Wenchao Jiang, Tom Rodden, Edwin Simpson, Steven Reece, Stephen Roberts, and Nicholas R. Jennings. 2016. A Disaster Response System based on Human-Agent Collectives. *Journal of Artificial Intelligence Research* 57 (2016), 661–708. <https://doi.org/10.1613/jair.5098>
- [10] Guillermo Ricardo Simari and Iyad Rahwan (Eds.). 2009. *Argumentation in Artificial Intelligence*. Springer.
- [11] David Smith. 2012. Planning as an Iterative Process. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*.