DOI:
10.1016/j.clsr.2020.105489

Document Version
Peer reviewed version

Building Trust in Human-Machine Partnerships*

Gerard Canal[a,*], Rita Borgo[a], Andrew Coles[a], Archie Drake[b], Dong Huynh[a], Perry Keller[b], Senka Krivić[a], Paul Luff[c], Quratul-ain Mahesar[a], Luc Moreau[a], Simon Parsons[d], Menisha Patel[c], Elizabeth I Sklar[e]

[a]*Department of Informatics, King's College London*
[b]*The Dickson Poon School of Law, King's College London*
[c]*King's Business School, King's College London*
[d]*School of Computer Science, University of Lincoln*
[e]*Lincoln Institute of Agrifood Technology, University of Lincoln*

Abstract

Artificial intelligence (AI) is bringing radical change to our lives. Fostering trust in this technology requires the technology to be transparent, and one route to transparency is to make the decisions that are reached by AIs *explainable* to the humans that interact with them. This paper lays out an exploratory approach to developing explainability and trust, describing the specific technologies that we are adopting, the social and organizational context in which we are working, and some of the challenges that we are addressing.

Keywords:
Artificial intelligence, explainable AI, planning.

1. Introduction

The full potential of artificial intelligence (AI) has yet to be realised, yet AI has already become a driving force in a radical transformation of human life. We posit that the influence of today's AI, in many ways, resembles the impact of the industrial revolution in the 18th and 19th centuries. In the industrial revolution, adaptations of law and regulation often ran well behind the negative effects of the new innovations of industrialisation. In

*Invited submission for Computer Law and Security Review following exAI2020
*Corresponding Author

today's AI revolution, there is broad consensus that the rapid development of AI technologies demands equally fast-paced and innovative development in governance policies, laws and regulations. If designed appropriately, many of the capabilities of AI systems will be enabled, while also preserving the personal and collective autonomy and dignity of the people who will benefit.

As part of that ambition, AI *explainability* is often seen as a key feature in the effort to achieve effective and also ethically and legally sound AI-driven decision making. *Explainability,* the ability of an AI system to explain its decisions or choices — made either by the AI system alone, or with a system that interfaces with the AI system — potentially offers the level of transparency that is necessary for people to trust in this decision making. Ideally, full explainability will lead to a cascade of trustworthiness that begins with AI developers and users, runs through to regulators and legislators, and finishes with the general public — who will know that, if necessary, any decisions made or recommended by an AI system can be explained.

At present, this notion of end-to-end technical, ethical and legal coherence in AI systems is a distant aspiration. AI technologies are developing quickly, outpacing changes in law and regulation. In that sphere, innovative principles, standards and guidance abound, but finding effective ways of operationalising these rights and duties is very much a work in progress. Data protection law, for example, is under-developed where its protections are most relevant, yet excessively burdensome where they are not.

Given the fact that the development of technology is outpacing legislation, other routes need to be identified for achieving the overarching goal of ensuring that AI systems are trustworthy for everyone significantly involved or affected. We argue that one way in which the linkage of explainability and trust can be built up is through *practical exploration.* That is to say, expectations about meaningful AI explainability shift according to different needs or purposes and what is required to foster trust necessarily follows those shifts. Each set of demands on an AI system, whether from the system's developers or from the ultimate beneficiaries of the system's decisions, requires an appropriate adaptation of the system's explanations in order to render its outputs trustworthy. In short, explainability and trust are multi-faceted, and each strand requires focused, context-driven attention.

The foundation for this exploratory approach to explainability and trust is a human-machine team, made up of one or more people and one or more AI systems. We argue that trust between the members of a human-machine team is best established when the humans involved are confident that the

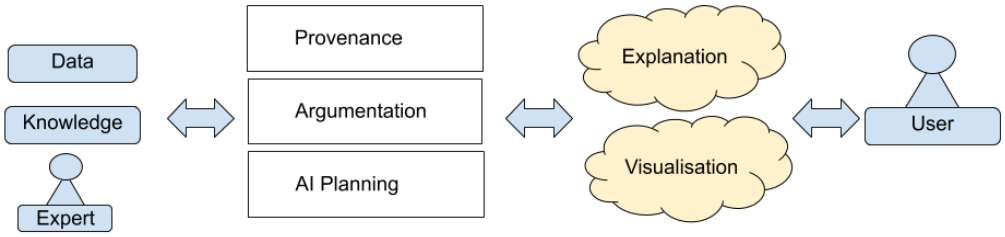

Figure 1: An overview of the elements we are developing.

following criteria hold:

- Criterion 1: All decisions are based on appropriate information and that information has been processed in suitable ways.
- Criterion 2: The reasons behind the decisions are communicated clearly and effectively.
- Criterion 3: The people impacted can engage in a process of discussing and questioning decisions with the AI system(s).

These broad criteria are useful in judging the soundness of explainability for any party affected by AI decision-making, and below we identify how we are working towards providing this ability. The nature and specificity of the information sought is, however, likely to vary according to the concerns of affected stakeholders, and, again, this is an aspect that we are addressing in our work.

Overall, we are concerned with the explainability that is sufficient to ensure that humans working with the machine are comfortable sharing responsibility. Without basic explainability of this kind in relation to the immediate purposes of the AI system, all other demands for explainability are premature. Plainly, while explainability and trust within the human-machine team is in this sense fundamental, it will often be ethically and legally defective, especially for AI systems that have significant effects on large numbers of citizens or consumers. Depending on the AI system and its purposes, adequate explainability may become highly complex as different expectations of trustworthiness are worked into its design and operation.

Our work focuses on the foundational explainability and trust of the human-machine team, bridging several different disciplines and involving relevant stakeholders from organisations in the commercial and charitable

sectors. Figure 1 broadly outlines the elements of our approach and the relationships amongst them. Our work is firmly positioned in the area of model-based, as opposed to data-driven AI[1], for which we have AI planning as a central component. Thus, our 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 corresponding 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.

Importantly, our work also extends to the organisational and social contexts in which explanation within the human-machine team is required, including the practical constraints affecting their production and the ways in which they are understood. In our view, a significant requirement for progress towards trusted and trustworthy AI is for research to build both on seminal work and lessons learned from "expert systems", as well as newer observations about the relationship of AI to social sciences [1].

The remainder of the paper is structured as follows: Section 2 describes the technical aspects that we consider in the pursuit of *computable explanations*, summarizing the different technical disciplines that we are bringing together to address this goal, along with various challenges we are facing. In Section 3, we illustrate how the social and organizational context shapes the requirements for explanations, and how we intend to evaluate the utility of the technical approaches developed. Finally, Section 4 concludes with a summary of salient points and a discussion of future directions.

2. Technical aspects

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

[1]That is, the AI is provided with a model of its task domain rather than extracting one from data. In terms of explainability, this is advantageous since the model itself can be used to build explanations.

highlight how these tools allow us to develop computational models for explanations and, ultimately, trust elicitation in human-machine teams, in particular how they address criteria 1, 2 and 3 raised above.

2.1. AI Planning

Planning is an essential component of any organization, work or autonomous system, and has been studied in AI since the 1970s [2]. With the recent development of industrial-strength approaches, automated planning has become a prominent technology. The problems to which planning can be applied span a wide range of areas, from optimisation to robotics, and may involve conflicting objectives and timing constraints.

The definition of 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. It does so by checking how to modify the initial state of the environment by applying some of the actions that are available, provided that some conditions for the actions are met. A solution to the problem is a plan detailing *what* actions to do, and *when* (in what order) to do them, in order to transform the initial state into the goal state.

Now, some stakeholders 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. As a result, their expectations may not mesh with the plan that the system comes up with, potentially creating mistrust. This may happen because of the stakeholder disagrees with the system about some aspects of the problem — for example about what actions are available — or it may be because of the complexity of the domain — the stakeholder may just not be able to compute a good solution themselves. Because decisions are based on a model, it is possible to showing the reasons behind each of the decisions, in terms of the aspects of the model that led to the decision. This is in contrast to systems that operate in a black box or data-based manner. Obtaining and presenting these reasons, criterion 2 from the introduction, can be used to build trust, as explored below.

However, this efficient communication of the reasons behind the planner decisions is a challenging task, and this is even more clear when applied to domain-independent technologies like planning. While domain-specific methods could be developed, generic alternatives may provide explanations independently of the underlying algorithms. We are currently tackling this problem by means of contrastive explanations [3, 4] where the planner used

to solve the problem is also employed to generate explanations. These explanations are presented as alternative plans including users' constraints, but more complex presentations involving natural language descriptions and visualizations will be also investigated (as described below).

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 [5], can be applied to planning systems, providing explanations according to criterion 1 from the introduction. 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 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 [6]. 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 [3], better understand decisions made by human-machine planning systems.

While provenance of data, plans, and plan executions appears to be very relevant to construct explanations, actual provenance for AI-based systems is likely to be very large; thus, the provenance elements that are most relevant for the purpose and for the recipient of the explanations need to be carefully selected and extracted. That selection is certainly non-trivial and would require exploration to develop an approach that instils trust in the generated explanations. Another concern is that provenance links up data, processes and methods, some of which are potentially private (e.g., intellectual property of an organisation, or confidential data about a data subject). Thus, a privacy-by-design methodology should be applied to explanation techniques for explanations to be trusted.

2.3. Argumentation

Argumentation [7] 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 three 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 [8]. This addresses criterion 2 from the introduction. Second, we have introduced argumentation schemes [9] that can be used to 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. This

helps to deepen the kind of reasons that can be provided, again addressing criterion 2. Third, we have demonstrated how the interpretation of a plan as an argument can be probed in a process of discussing and questioning the plan, addressing criterion 3 from the introduction.

The major challenge here is generating the kinds of explanation and probing dialogue that human users will require from the system. From previous work [10], we know that providing argumentation-based explanations can improve the solutions found by some kinds of human-machine team. The question is whether the kinds of arguments that we can construct to explain plans are considered adequate. We will be carrying out experiments with human participants to establish whether this is the case, and, if not, how we need to modify the elements that we have developed.

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, which addresses criterion 3 from the introduction.

As per criterion 2, 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 [3] — providing users counterfactual examples of plans considering users questions; planning justifications — providing reasons for changes in the planning process; ethical comparison of plans [12]; argumentation based explanations — providing arguments expressed in natural language [9]; 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(s) and stakeholders or as an exploration tool to dive inside its decision flow. Explanation visualization addresses criterion 2 from the introduction. Several attempts have been made to develop platforms to support the latter [13] but progress is still needed on the former. In [14] we made a first attempt to enhance the traditional interface based

communication approach to enrich the conversation flow and embed stakeholders 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 [15] 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 capabilites as well as level of expertise.

3. Social and organizational aspects

As well as the technical challenges of extracting range of different forms of content for an explanation of plan, summarising this in a suitable way and then finding a way to present the outcomes, there are considerable difficulties in producing explanations that are useful and appropriate for users of different kinds. Even with an explanation that contains appropriate content, including explanations of the reasoning of the planning systems that include details of the provenance that are presented in some kind of argument and, perhaps, visualised in a clear manner, it will not necessarily be the case that the explanation will be useful or appropriate for a user, let alone trustworthy. Plans, of the kinds that can be developed through the AI systems, are complex and are most likely to be used in specialised domains where there are significant difficulties in planning and scheduling activities. Therefore, we need to understand the needs of users in organisations, their requirements, so that we can propose how the explanations can be designed.

To provide a concrete motivating examples, we are collaborating with an organisation which already uses AI, and in particular, planning systems for very specific tasks and undertaking a study involving the different kinds of people involved in the planning process. We draw from qualitative social

scientific approaches, undertaking detailed semi-structured interviews with experts and potential users to elicit their experience and perceptions regarding planning and explanations. We consider not only the needs of those who use plans in their work, but those who produce them. We seek to understand better the nature of explanations in organisational contexts and the practical constraints in which they are produced and understood. Hence, we are investigating 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. Although we are still in the process of analysing the interviews, it is apparent that what constitutes an explanation is closely tied to particular features of the context; participants design their explanations for particular uses (and particular colleagues). They typically are concise and shaped to the specific circumstances at hand. Also, explanations are not produced as distinct and separable items, but are part of an interaction, and are shaped according to the demands of that interaction. It seems that this tailorability, the ways in which explanations are shaped for a particular recipient help make them to be trusted.

The studies with participants in organisations will also inform how we assess the techniques and models we develop in the project. As it will be infeasible to deploy some of the prototype approaches we plan to undertake quasi-naturalistic studies of these technologies, where we can investigate how users understand explanations in realistic, but more general, activity. As an example, we are developing a robotic scenario where a robot acts in an office setting and look for cases for explanations there. Such a scenario has the advantage that it may be more relatable for the general public. The studies will allow us to assess different forms of explanation and also whether and how the explanations can be considered trustworthy. As this is a complex matter to investigate, the experiments offer the potential for exposing issues associated with trust by breaching this trust. By developing so-called breaching experiments [16]. We can carefully explore some of the ways that explanations might not be trusted, and hence inform approaches for making them more trustworthy.

4. Summary and discussion

Our work so far suggests that there are several forms of explanation for AI planning that are computable. As in [3], we can produce contrastive explanations, and we can also generate explanations that take a more causal direction [9]. 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 — and to whom, in which contexts, and how — is ongoing work.

We have centered our work following three criteria: decisions based on appropriate information, clear and effective communication of the reasons behind the AI decisions, and engagement of users in discussions with the system. We have argued how Provenance, AI Planning, Argumentation, and Visualization can be joined together to address these three criteria. We do this by tracking the data flow involved in the decisions with Provenance, computing a solution using Planning, and explaining those by means of Argumentation. The resulting explanations are then effectively presented to the user in the form of explanations and visualizations.

The effectiveness of the explanations that we can currently compute will be tested through studies with human participants, 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, how these might be satisfied, and how they fit into existing legal and regulatory frameworks [17]. Whatever the answers in terms of the computability of explanations, clearly there are limits. We expect a better understanding of the limits of computational approaches to explanation to be useful to other disciplines, notably law, in framing alternative approaches to the problem of trust in human-machine interactions.

In our opinion, this is a critical moment for research to make genuinely impactful contributions based on XAI across several dimensions. Together with associated legislation the European General Data Protection Regulation 2016, with its so-called "right to explanation", has drawn attention to XAI as a key concern not just for software engineers but for any organisation

that uses or wants to use AI in its work [18]. The race is on for businesses across sectors to gain a competitive edge using XAI capability to improve the productivity of staff and the quality of interactions with consumers. And it is becoming clearer that thinking on XAI can inform policymakers looking for ways to anticipate and mitigate some of the negative implications of current economic transformations for individuals and communities.

Acknowledgement

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

References

[1] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38.

[2] R. E. Fikes, N. J. Nilsson, Strips: A new approach to the application of theorem proving to problem solving, Artificial intelligence 2 (3-4) (1971) 189–208.

[3] M. Cashmore, A. Collins, B. Krarup, S. Krivic, D. Magazzeni, D. Smith, Towards Explainable Planning as a Service, in: ICAPS-19 Workshop on Explainable Planning, 2019.

[4] A. Collins, D. Magazzeni, S. Parsons, Towards an argumentation-based approach to explainable planning, in: ICAPS-19 Workshop on Explainable Planning, 2019.

[5] L. Moreau, P. Missier, PROV-DM: The PROV data model, Tech. rep., World Wide Web Consortium, W3C Recommendation (2013).
URL `http://www.w3.org/TR/2013/REC-prov-dm-20130430/`

[6] S. D. Ramchurn, T. D. Huynh, F. Wu, Y. Ikuno, J. Flann, L. Moreau, J. E. Fischer, W. Jiang, T. Rodden, E. Simpson, S. Reece, S. Roberts, N. R. Jennings, A disaster response system based on human-agent collectives, Journal of Artificial Intelligence Research 57 (2016) 661–708. doi:10.1613/jair.5098.

[7] G. R. Simari, I. Rahwan (Eds.), Argumentation in Artificial Intelligence, Springer, 2009.

[8] X. Fan, On generating explainable plans with assumption-based argumentation, in: International Conference on Principles and Practice of Multi-Agent Systems, Springer, 2018, pp. 344–361.

[9] Q. Mahesar, S. Parsons, Argument schemes for explainable planning (2020). arXiv:2005.05849.

[10] M. Q. Azhar, E. I. Sklar, A study measuring the impact of shared decision making in a human-robot team, The International Journal of Robotics Research 36 (5-7) (2017) 461–482.

[11] D. Smith, Planning as an iterative process, in: Proceedings of the 26th AAAI Conference on Artificial Intelligence, 2012.

[12] B. Krarup, S. Krivic, F. Lindner, D. Long, Towards Contrastive Explanations for Comparing the Ethics of Plans, in: ICRA-20 Workshop on Against Robot Dystopias, 2020.

[13] A. Chatzimparmpas, R. M. Martins, I. Jusufi, A. Kerren, A survey of surveys on the use of visualization for interpreting machine learning models, Information Visualization.

[14] R. Borgo, M. Cashmore, D. Magazzeni, Towards providing explanations for AI planner decisions, CoRR abs/1810.06338 (2018).

[15] R. Borgo, J. Kehrer, D. Chung, E. Maguire, R. Laramee, H. Hauser, M. Ward, M. Chen, Glyph-based visualization: Foundations, design guidelines, techniques and applications, 2013.

[16] H. Garfinkel, Studies in ethnomethodology, Englewood Cliffs, NJ: Prentice-Hall, 1967.

[17] P. Keller, A. Drake, Exclusivity and Paternalism in the public governance of explainable AI, Computer Law and Security Review (2020).

[18] The Information Commissioner's Office (ICO), Explaining decisions made with ai, `https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-ai`, Accessed: August 2020.