Towards Scrutable Autonomous Systems

Judith Masthoff, Nir Oren, Kees van Deemter, Wamberto W. Vasconcelos
Department of Computing Science, University of Aberdeen, AB24 3UE, UK
{j.masthoff, n.oren, k.vdeemter, w.w.vasconcelos}@abdn.ac.uk

Introduction. Distributed autonomous systems (DAS) suffer from opacity: it is difficult for humans to establish why certain behaviours occurred, and what alternatives existed. Human performance problems with autonomous systems are exacerbated by the lack of information on how automation is working, poor feedback about the automation's activities, absence of support for operators to communicate with automation, and inadequate explanation of its reasoning processes [1]. The recently EPSRC-funded SAsSy project proposes to investigate computational mechanisms for supporting the scrutiny of coordination activities in DAS, providing transparency, whereby the internal workings of DAS are exposed to humans. The project will investigate i) what information is needed to support the scrutiny of DAS, ii) mechanisms to gather this information, and iii) how best to present this information to humans, so as to increase their understanding of and trust in the DAS.

We propose to utilise formal argumentation techniques [2] in distributed planning, because they allow generating and gathering relevant information about the components' internal workings, their interactions and the rationale for joint decision making, thereby providing the basic constructs needed in the explanation process. However, individual arguments can be large and complex, and many arguments may be associated with a joint plan. To make formal arguments useful to humans, we will combine argumentation with Information Presentation (IP) techniques for expressing information in a human-understandable format and simplifying, summarising and aggregating information as needed. The IP will be guided by extensive knowledge acquisition, user modelling and evaluation with business partners, to base the choice of IP on. Ultimately, we aim at raising awareness of the importance of scrutability of DAS in general.

Shortcomings of related work. Existing frameworks for evaluating agent-oriented methodologies (e.g., [3]), typically list many criteria, which tend to be unrelated to end-users' needs. Some of these criteria which appear related to end-users, such as understandability, analysability, accessibility, required expertise, and usability, are in fact aimed at developers. For example, usability has been defined as the number of public attributes in an agent class. Nevertheless, it has been acknowledged that humans may play an important part in DAS, in particular managing multi-agent teams or participating in hybrid teams [4]. Many scenarios need to keep the human in the loop with useful information and to provide efficient methods of interacting with the DAS. Some evaluations of DAS with users have been reported (e.g., [4]), but we have not found evaluations of scrutability *per se*. Within the wider intelligent systems community, there has been an increased interest in more user-centred evaluation metrics, including scrutability. Within our envisaged research, scrutability is a measure of whether users understand a plan and the motivation behind it. Explanations have been investigated for expert systems providing decision support and more recently for recommender systems. However, these systems use explanations and reasoning techniques (e.g., Bayesian models, collaborative filtering) very different from the ones required in distributed planning. Moreover, existing work does not deal with multiple agents. Evaluations of explanations in expert systems have focused largely on users' acceptance of the system as a whole or acceptance of the system's decisions, rather than on scrutability. There is existing work on supporting developers to debug DAS using graphical representations of agent interaction protocols, and on real-time dialogues between humans and individual agents. Contrasting with this existing work, we will build a scrutable DAS that provides end-users with a global perspective, enabling inspection of the DAS as a whole, and helping to increase end-user trust in the DAS. The project will be user-driven, investigating through knowledge acquisition what questions users want to ask, and empirically evaluated.

Challenges. The project contains difficult challenges in each of the research areas that are involved. Especially relevant to this forum, and one of the main IP challenges, is to find optimal English wordings to paraphrase potentially complex logical formulas. In fact, this is a famous open problem in Natural Language Generation (NLG), known as the Logical Form Equivalence problem [5]. Given an input formula F (which is a formula to be scrutinised by the user) "translating" F directly into English might lead to an unnecessarily complex sentence. The problem, in this case, is to first convert F into a simpler but logically equivalent formula F′, which needs to be found first. This is challenging both computationally (particularly if the logic is very expressive) and empirically (because we do not know in advance which English sentences are easiest for the reader to understand). The research thus needs to combine techniques from a number of specialities (including planning, NLG and psycholinguistics) in a novel way.

Hypotheses and Goals: We hypothesise that i) formal argumentation techniques can be used to improve distributed planning (by obtaining plans more quickly); ii) formal argumentation techniques can gather the information required to support the scrutiny of DAS; iii) multimodal presentations of the information allow non-experts to scrutinise the DAS. In order to evaluate these hypotheses we shall i) develop formal representations and associated mechanisms for arguments, plans, preferences and norms and combine these with existing planning mechanisms; ii) propose dialogues for joint decision-making about plans; and iii) adapt and extend IP techniques to allow human scrutiny of large amounts of complex information about distributed planning.

References

1. O'Hara, J and Higgins, J. *Human-system interfaces to automatic systems: Review guidance and technical basis*. Energy Sciences and Technology Department, Brookhaven National Laboratory, Upton, New York, USA. BNL-91017-2010. http://www.bnl.gov/isd/documents/71082.pdf. 2010.
2. Dung, P. M. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, **77**(2):321–357. 1995.
3. Garcia, E., Giret, A., and Botti, V. Evaluating software engineering techniques for developing complex systems with multiagent approaches. *Information and Software Technology*, 53(5):494–506. 2011.
4. Sycara, K., Norman, T. J., Giampapa, J. A., Kollingbaum, M. J., Burnett, C., Masato, D., McCallum, M. and Strub, M. H. Agent support for policy-driven collaborative mission planning. *The Computer Journal*, 53(5):528-540. 2010.
5. Shieber, S. M. The problem of logical form equivalence. *Comp. Ling.*, 19(1):179– 190. 1993.