Evaluating a Mechanism for Explaining BDI Agent Behaviour
Extended Abstract

Michael Winikoff
Victoria University of Wellington
Wellington, New Zealand
michael.winikoff@vuw.ac.nz

Galina Sidorenko
Halmstad University
Halmstad, Sweden
galina.sidorenko@hh.se

ABSTRACT
We conducted a survey to evaluate a previously proposed mechanism for explaining Belief-Desire-Intention (BDI) agents using folk psychological concepts (belief, desires, and valuings). We also consider the relationship between trust in the specific autonomous system, and general trust in technology. We find that explanations that include valuings are particularly likely to be preferred by the study participants. We also found evidence that single-factor explanations, as used in some previous work, are too short.

KEYWORDS
Explanation; Explainable Agency; Belief-Desire-Intention (BDI)

2 METHODOLOGY
We use the following scenario: Imagine that you have a smart phone with a new smart software assistant, SAM. Unlike current generations of assistants, this one is able to act proactively and autonomously to support you. SAM knows that usually you use one of the following three options to get home: Walking, Cycling (if a bicycle is available), and Catching a bus (if money is available). You are about to leave to go home, when the phone alerts you that SAM has just bought you a ticket to catch the bus home. This surprises you, since you typically walk or cycle home. You therefore push the "please explain" button.

Our explanations are formed out of four elements: desires (D), beliefs (B), valuings (V), and links (L). For example: A bicycle was not available (B), money was available (B), the made choice (catch bus) has the shortest duration to get home (in comparison with walking) and I believe that is the most important factor for you (V), I needed to buy a bus ticket in order to allow you to go by bus (L), and I desire to allow you to catch the bus (D).

Each participant was presented with five possible explanations in random order. Explanation E1 includes all four elements, E2 filters out the desires and links, E3 includes only valuings, E4 includes only beliefs, and E5 includes only beliefs and desires.

For each of the five explanations E1-E5 participants were asked to indicate on a Likert scale of 1-7 (1 = “Strongly Disagree”, 7 = “Strongly Agree”) how much they agree or disagree with the following statements: “This explanation is Believable (i.e. I can imagine a human giving this answer)”, “This explanation is Acceptable (i.e. this is a valid explanation of the software’s behaviour)”, and “This explanation is Comprehensible (i.e. I understand this explanation)”.

Participants were then asked to rank the explanations from most to least preferred. They were also asked to indicate the extent to which they agreed with the statement “I trust SAM because it can provide me a relevant explanation for its actions” (7 point Likert scale). Next, the survey asked a number of questions to assess and obtain information about general trust in technology, including attitude to Artificial Intelligence. The 11 questions consisted of 7 questions that were adopted from McKnight et al. [17, Appendix B]. Specifically, we used the four questions that McKnight et al. used to assess faith in general technology (item 6 in their appendix), and the three questions that they used to assess trusting stance (general technology, item 7). We also had four questions that assessed attitudes towards Artificial Intelligence. Finally, the respondents were asked to provide demographic information. The next section summarises key results, for full details see [27].
3 KEY RESULTS (SEE [27] FOR DETAILS)

Believability, Acceptability, and Comprehensibility of Explanations: Figure 1 depicts the statistically significant differences on the first set of questions. We found that overall E2 can be seen as the best explanation since it is ranked statistically significantly differently to all other explanations (with a higher median) on at least one of the three characteristics (Believability, Acceptability, and Comprehensibility), but no other explanation is better than it on any characteristic. Next are E1 and E3 which are statistically different (specifically better) than E4 and E5 on some characteristics (for E1 Comprehensibility and Acceptability but not Believability, and for E2 Believability and Acceptability, but not Comprehensibility).

To analyse the ranked data (“rankings of explanations” and “effects of explanation components”) we employed a general discrete choice model (linear mixed model), using a ranked-ordered logit model which is also known as an exploded logit [1]. A discrete choice model is a general and powerful technique for analysing which factors contributed to the outcome of a made choice. It is required in this case because each of the five explanations being ranked represented a combination of explanatory factor types. The ranked-ordered logit is used to deal with the fact that the data represents a ranking: after selecting the most preferred explanation, the next selection is made out of the remaining four explanations. This means that the selections are not independent.

Rankings of Explanations: We found that E2 is most preferred, followed by E1 and E3, which are not significantly differently ranked, and then E4 and E5 (also not statistically significantly different in ranking). In other words, we have three tiers: E2 (most preferred), E1 and E3 (less preferred than E2), and E4 and E5 (least preferred). This is consistent with the results shown in Figure 1.

Effects of Explanation Components: We found that explanations containing valuations are much more likely to be preferred over explanations where valuations are absent (1002.3% increase in the odds), and that the presence of beliefs (respectively desires) also make an explanation more likely to be preferred (127% increase for beliefs, 71.6% for desires). On the other hand, the presence of a link explanation reduces the likelihood of preference by 68.65%. The difference between preferring B and D is not statistically significant, whereas the difference among all others components is significant. This analysis shows that of the four factors that are included in the explanations, the presence of V components most strongly (and significantly) correlates with higher preference for the explanation.

Relationship between trust in SAM and broader trust: We found a clearly significant ($p = 3.85 \times 10^{-5}$) positive but moderate correlation ($\rho_{S} = 0.46$) between the trust in SAM and general trust ($\rho_{S} = 0.46$, $n = 74$, $p = 3.8 \times 10^{-5}$). Thus, high values of background trust in technology are associated with high “trust in SAM” scores. Since our survey assessed trust in technology before participants were introduced to SAM, we have that trust in technology cannot be influenced by anything related to SAM. Therefore, the correlation can be interpreted as indicating that while trust in technology in general (including AI) influences trust in SAM (as might be expected), it does not determine it. This is an encouraging finding: if we had found that preexisting trust in technology and AI in general strongly affected (or even determined) trust in a given autonomous system, then there would be a limited (or no) role for explanations to affect the level of trust.

4 DISCUSSION

Based on our findings, we provide the following advice to guide the further development of explanation mechanisms for autonomous agents.

Firstly, it is clear that valuations are valued, which is consistent with the findings of the previous evaluation [26]. Therefore valuations should be included in explanations.

Secondly, we found that explanations including link components were less likely to be preferred. The evaluation by Harbers et al. [11] also found that link explanations were barely selected as preferred. However, we only had one explanation that included links (E1), and it may also be that the lower preference for this explanation reflects its length. We therefore do not recommend excluding link explanatory components at this point, but rather suggest that further evaluation would help to clarify whether they are indeed seen as less preferred.

Thirdly, we did not find that users prefer short explanations. The most preferred explanation was E2, which is longer than E3 and E4. On the other hand, the longest explanation (E1) was not the least preferred. Although the length of an explanation clearly can play a role, with too-long explanations being less useful, our findings do not support the approach taken by previous work (e.g. [4, 10, 11, 13, 15]) to limit explanations to a single explanatory element (e.g. a single belief or a single desire).

There is scope for further evaluation, with different scenarios, and with different forms of explanations. Two specific forms of explanation that would be good to consider are emotions, and interactive explanations. Keptein et al. [14] argue that explanations should include emotions. We only considered explanations that were presented to the user all at once. However, explanations can also be presented in the form of a dialogue, with an initial reason being given, and then additional information being provided as the user interacts with the system (See e.g. [7, 8, 22, 25]).
ACKNOWLEDGMENTS

We would like to thank Dr. Damien Mather, at the University of Otago, for statistical advice. This work was supported by a University of Otago Research Grant (UORG).

REFERENCES


doi:10.1007/978-3-642-16178-0_5

doi:10.1007/978-3-030-17294-7_3


