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)

ACM Reference Format:
Michael Winikoff and Galina Sidorenko. 2023. Evaluating a Mechanism for Explaining BDI Agent Behaviour: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

It is now widely accepted that explainability is crucial for supporting an appropriate level of trust in autonomous and intelligent systems (e.g. [9, 12, 21]), and for other reasons (e.g. making systems understandable [24], accountable [5], challengeable, predictable, verifiable, and traceable [24]). We focus on *autonomous agents*, which includes a wide range of systems both embodied (e.g. robots) and non-embodied (e.g. smart personal assistants) [18, 19, 23].

Prior work has shown that humans use the concepts of beliefs, desires, and valuings[1] when explaining their behaviour [16], and subsequently we proposed [28] a mechanism that allows Belief-Desire-Intention (BDI) agents [2, 3, 20] (augmented with a representation for valuings, following [6]) to provide explanations of their actions in terms of these concepts.

In this paper we conduct an empirical human subject evaluation[2] of this mechanism, answering the questions: "What forms of explanation of autonomous agents are preferred?" and "to what extent is trust in a given system determined by a person's more general attitudes towards technology, and towards Artificial Intelligence?".

[1]Defined as things that "*directly indicate the positive or negative affect toward the action or its outcome*"
[2]This paper differs from an earlier evaluation [26]: it includes links, considers general trust in technology, and conducts a deeper more sophisticated analysis.

The authors were at the University of Otago, New Zealand, when most of the work was done.

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

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[3] 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]Ethics approval was given by University of Otago (Category B, D18/231). Participants were recruited by advertising in undergraduate lectures at the Otago Business School, by email to students at institutions of Frank and Virginia Dignum, and by posting on social media. The survey is available at: https://www.dropbox.com/s/ec6fg3u1rqhytcb/Trust-Autonomous-Survey.pdf.

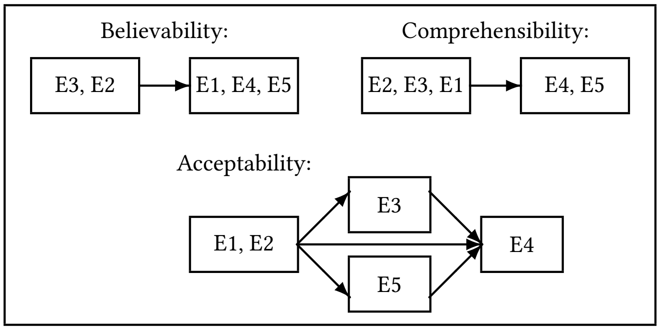

Figure 1: Visual representation of the significance results. An arrow indicates a statistically significant difference (arrow is directional from better to worse)

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[4]. 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 valuings are much more likely to be preferred over explanations where valuings are absent (1002.3% increase in the odds), and that the presence of beliefs (respectively desires) also make an explanation more like 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 ($\rho_S = 0.46$, $n = 74$, $p = 3.8\times10^{-5}$). 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 valuings are valued, which is consistent with the findings of the previous evaluation [26]. Therefore valuings 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]).

[4] We used paired Wilcoxon-signed rank, using a significance level of 0.005 rather than 0.05 to avoid type II errors, given the number of tests performed. The significance level is calculated as $\sqrt[10]{0.95} = 0.9948838$, giving a threshold for significance of around 0.005.

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

[1] Paul D. Allison and Nicholas A. Christakis. 1994. Logit Models for Sets of Ranked Items. *Sociological Methodology* 24 (1994), 199–228. https://www.jstor.org/stable/270983

[2] Michael E. Bratman. 1987. *Intentions, Plans, and Practical Reason.* Harvard University Press, Cambridge, MA.

[3] M. E. Bratman, D. J. Israel, and M. E. Pollack. 1988. Plans and resource-bounded practical reasoning. *Computational Intelligence* 4 (1988), 349–355.

[4] Joost Broekens, Maaike Harbers, Koen V. Hindriks, Karel van den Bosch, Catholijn M. Jonker, and John-Jules Ch. Meyer. 2010. Do You Get It? User-Evaluated Explainable BDI Agents. In *Proceedings of the 8th German Conference on Multiagent System Technologies (MATES 2010) (LNCS, Vol. 6251)*, Jürgen Dix and Cees Witteveen (Eds.). Springer, 28–39. https://doi.org/10.1007/978-3-642-16178-0_5 doi:10.1007/978-3-642-16178-0_5.

[5] Stephen Cranefield, Nir Oren, and Wamberto Weber Vasconcelos. 2018. Accountability for Practical Reasoning Agents. In *Agreement Technologies - 6th International Conference, AT 2018, Bergen, Norway, December 6-7, 2018, Revised Selected Papers (Lecture Notes in Computer Science, Vol. 11327)*, Marin Lujak (Ed.). Springer, 33–48. https://doi.org/10.1007/978-3-030-17294-7_3

[6] Stephen Cranefield, Michael Winikoff, Virginia Dignum, and Frank Dignum. 2017. No Pizza for You: Value-based Plan Selection in BDI Agents. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. 178–184. https://doi.org/10.24963/ijcai.2017/26

[7] Louise A. Dennis and Nir Oren. 2021. Explaining BDI Agent Behaviour through Dialogue. In *AAMAS '21: 20th International Conference on Autonomous Agents and Multiagent Systems, Virtual Event, United Kingdom, May 3-7, 2021*, Frank Dignum, Alessio Lomuscio, Ulle Endriss, and Ann Nowé (Eds.). ACM, 429–437. https://doi.org/10.5555/3463952.3464007

[8] Louise A. Dennis and Nir Oren. 2022. Explaining BDI agent behaviour through dialogue. *Auton. Agents Multi Agent Syst.* 36, 1 (2022), 29. https://doi.org/10.1007/s10458-022-09556-8

[9] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena. 2018. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines* 28, 4 (01 Dec 2018), 689–707. https://doi.org/10.1007/s11023-018-9482-5

[10] Maaike Harbers. 2011. *Explaining Agent Behavior in Virtual Training.* SIKS Dissertation Series No. 2011-35. SIKS (Dutch Research School for Information and Knowledge Systems).

[11] Maaike Harbers, Karel van den Bosch, and John-Jules Ch. Meyer. 2010. Design and Evaluation of Explainable BDI Agents. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Intelligent Agent Technology, IAT 2010, Toronto, Canada, August 31 - September 3, 2010*, Jimmy Xiangji Huang, Ali A. Ghorbani, Mohand-Said Hacid, and Takahira Yamaguchi (Eds.). IEEE Computer Society Press, 125–132. https://doi.org/10.1109/WI-IAT.2010.115

[12] High-Level Expert Group on Artificial Intelligence. 2020. The Assessment List for Trustworthy Artificial Intelligence. https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

[13] Frank Kaptein, Joost Broekens, Koen V. Hindriks, and Mark A. Neerincx. 2017. Personalised self-explanation by robots: The role of goals versus beliefs in robot-action explanation for children and adults. In *26th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2017, Lisbon, Portugal, August 28 - Sept. 1, 2017*. IEEE, 676–682. https://doi.org/10.1109/ROMAN.2017.8172376

[14] Frank Kaptein, Joost Broekens, Koen V. Hindriks, and Mark A. Neerincx. 2017. The role of emotion in self-explanations by cognitive agents. In *Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, ACII Workshops 2017, San Antonio, TX, USA, October 23-26, 2017*. IEEE Computer Society, 88–93. https://doi.org/10.1109/ACIIW.2017.8272595

[15] Frank Kaptein, Joost Broekens, Koen V. Hindriks, and Mark A. Neerincx. 2019. Evaluating Cognitive and Affective Intelligent Agent Explanations in a Long-Term Health-Support Application for Children with Type 1 Diabetes. In *8th International Conference on Affective Computing and Intelligent Interaction, ACII 2019, Cambridge, United Kingdom, September 3-6, 2019*. IEEE, 1–7. https://doi.org/10.1109/ACII.2019.8925526

[16] Bertram F. Malle. 2004. *How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction.* The MIT Press. ISBN 0-262-13445-4.

[17] D. Harrison Mcknight, Michelle Carter, Jason Bennett Thatcher, and Paul F. Clay. 2011. Trust in a Specific Technology: An Investigation of Its Components and Measures. *ACM Transactions on Management Information Systems* 2, 2, Article 12 (July 2011), 25 pages. https://doi.org/10.1145/1985347.1985353

[18] J. Müller and K. Fischer. 2014. Application Impact of Multi-agent Systems and Technologies: A Survey. In *Agent-Oriented Software Engineering*, O. Shehory and A. Sturm (Eds.). Springer Berlin Heidelberg, 27–53. https://doi.org/10.1007/978-3-642-54432-3_3

[19] S. Munroe, T. Miller, R.A. Belecheanu, M. Pechoucek, P. McBurney, and M. Luck. 2006. Crossing the Agent Technology Chasm: Experiences and Challenges in Commercial Applications of Agents. *Knowledge Engineering Review* 21, 4 (2006), 345–392.

[20] Anand S. Rao and Michael P. Georgeff. 1992. An Abstract Architecture for Rational Agents. In *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning*, C. Rich, W. Swartout, and B. Nebel (Eds.). Morgan Kaufmann Publishers, San Mateo, CA, 439–449.

[21] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M. Howard, and Alan R. Wagner. 2016. Overtrust of Robots in Emergency Evacuation Scenarios. In *The Eleventh ACM/IEEE International Conference on Human Robot Interation, HRI 2016, Christchurch, New Zealand, March 7-10, 2016*, Christoph Bartneck, Yukie Nagai, Ana Paiva, and Selma Sabanovic (Eds.). IEEE/ACM, 101–108. https://doi.org/10.1109/HRI.2016.7451740

[22] Elizabeth I. Sklar and Mohammad Q. Azhar. 2018. Explanation through Argumentation. In *Proceedings of the 6th International Conference on Human-Agent Interaction, HAI 2018, Southampton, United Kingdom, December 15-18, 2018*, Michita Imai, Tim Norman, Elizabeth Sklar, and Takanori Komatsu (Eds.). ACM, 277–285. https://doi.org/10.1145/3284432.3284470

[23] M. Birna van Riemsdijk, Catholijn M. Jonker, and Victor R. Lesser. 2015. Creating Socially Adaptive Electronic Partners: Interaction, Reasoning and Ethical Challenges. In *Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, Gerhard Weiss, Pinar Yolum, Rafael H. Bordini, and Edith Elkind (Eds.). ACM, 1201–1206. http://dl.acm.org/citation.cfm?id=2773303

[24] Ruben S. Verhagen, Mark A. Neerincx, and Myrthe L. Tielman. 2021. A Two-Dimensional Explanation Framework to Classify AI as Incomprehensible, Interpretable, or Understandable. In *Explainable and Transparent AI and Multi-Agent Systems - Third International Workshop, EXTRAAMAS 2021, Virtual Event, May 3-7, 2021, Revised Selected Papers (Lecture Notes in Computer Science, Vol. 12688)*, Davide Calvaresi, Amro Najjar, Michael Winikoff, and Kary Främling (Eds.). Springer, 119–138. https://doi.org/10.1007/978-3-030-82017-6_8

[25] Michael Winikoff. 2017. Debugging Agent Programs with "Why?" Questions. In *Proceedings of the 16th Conference on Autonomous Agents and Multiagent Systems*. 251–259.

[26] Michael Winikoff, Virginia Dignum, and Frank Dignum. 2018. Why Bad Coffee? Explaining Agent Plans with Valuings. In *Computer Safety, Reliability, and Security - SAFECOMP 2018 Workshops, ASSURE, DECSoS, SASSUR, STRIVE, and WAISE, Västerås, Sweden, September 18, 2018, Proceedings (Lecture Notes in Computer Science, Vol. 11094)*, Barbara Gallina, Amund Skavhaug, Erwin Schoitsch, and Friedemann Bitsch (Eds.). Springer, 521–534. https://doi.org/10.1007/978-3-319-99229-7_47

[27] Michael Winikoff and Galina Sidorenko. 2021. Evaluating a mechanism for explaining BDI agent behaviour. https://profwinikoff.files.wordpress.com/2021/07/evaluating.pdf.

[28] Michael Winikoff, Galina Sidorenko, Virginia Dignum, and Frank Dignum. 2021. Why bad coffee? Explaining BDI agent behaviour with valuings. *Artificial Intelligence* 300 (2021), 31. https://doi.org/10.1016/j.artint.2021.103554