Interactive Explanations to Resolve Misalignments in Behaviour Support Agents

Johanna Wolff[1,*], Victor de Boer[2], Dirk Heylen[1] and M. Birna van Riemsdijk[1]

[1]University of Twente, Drienerlolaan 5, 7522 NB Enschede, The Netherlands
[2]Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands

Abstract

While using a behaviour support agent, a user's requirements and the situation they are in may change. This can lead to misalignments between the agent's recommendations and the user's wants and needs. In order to resolve these, the agent and the user need to understand each other's reasoning process. We introduce the structure of an agent using knowledge-based reasoning which can be directly explained and a user model which can by updated. We then propose a framework for a human-agent dialogue which is designed to identify the cause of a misalignment within the agent's reasoning and elicit the necessary information from the user to update the agent's knowledge base and realign the agent and the user.

Keywords

Behaviour Support Agent, Explainable AI, Misalignment Scenarios

1. Introduction

Making effective behaviour support agents [1] is a significant area of interest within the field of artificial intelligence, especially within hybrid intelligence [2]. These agents are intended to support behaviour in a personalised way, over a long period of time. To achieve this, the user has to be able to communicate their wants and needs to the agent to adapt the recommendations [3] and resolve potential misalignments [4]. Misalignments can be caused by a variety of issues and resolving them can be complicated; the user and the agent may not understand each other's reasoning. If an agent recommends going for a run and the user states "I don't want to go running because it is raining", this can manifest in the agent's reasoning in different ways. For example, the context "it is raining" may not have been detected, running in the rain should be avoided if possible or running should not be considered at all when it is raining. For effective realignment, the correct cause must be identified and the agent's knowledge base adjusted accordingly. This requires the user and the agent to have a shared mental model of the situation [5]. During the realignment process, there must be explanations for the agent's reasoning as well as space for the user to provide the necessary information to update the agent.

Knowledge-based methods can be used to explicitly represent the reasoning of the agent. This lets the agent describe how it reaches its conclusions when designing explanations and update the knowledge base of the agent if the user determines that information is incorrect or missing [6]. We base our interaction on an agent whose reasoning uses Default Logic [7], as it provides a way to reason with assumptions that are normally true but might have exceptions. Additionally, we can include conflicting motivations and possibilities, which can be can resolved using a preference ordering over the possible outcomes of the reasoning. For example, the user may usually want to go running on Tuesdays but not if it is raining or if they have already been running on Monday. In these cases the user prefers to do a yoga session, which is always possible but not preferred.

While these methods are inherently explainable, as a proof can be given for each conclusion, a proof

Hybrid Human Artificial Intelligence (HHAI) 2025, June 09–13, 2025, Pisa, Italy
*Corresponding author.
j.d.wolff@utwente.nl (J. Wolff); v.de.boer@vu.nl (V. d. Boer); d.k.j.heylen@utwente.nl (D. Heylen); m.b.vanriemsdijk@utwente.nl (M. B. v. Riemsdijk)
0009-0005-0178-9570 (J. Wolff); 0000-0001-9079-039X (V. d. Boer); 0000-0003-4288-3334 (D. Heylen); 0000-0001-9089-5271 (M. B. v. Riemsdijk)

generally does not qualify as an explanation [8], since it is not understandable to users without prior knowledge of formal proof methods. To mitigate this, logical inference rules can be translated into natural language and presented as a relation between the preconditions and consequences of these rules [9]. By breaking the reasoning down into individual steps and letting the user ask for explanations with different levels of complexity, the information can be presented in a way that most users can understand [10, 11]. When designing explanations, it is also important to consider the context and the goal of the explanation [12]. Our goal is to determine the cause of a misalignment between the agent and the user. This means that we are not explaining the agent's reasoning with the objective of convincing the user or trying to gain their trust but rather to allow the user to scrutinise the agent [13].

In this paper we describe how we designed such a dialogue between an agent and a user so that both sides are able to explain and understand where a potential misalignment originates from and then resolve this. We begin by giving an overview of the structure of the agent's knowledge base and reasoning process and then explain the dialogue flow between the agent and the user that we have designed. Finally we discuss the Co-12 properties [14] of our explanations and future work.

2. Structure of the Agent

We begin by introducing the basic agent structure that our interaction will be based on. To provide behaviour support, the agent represents information about the context, the goals the user may pursue and the actions that can be taken to reach these. In behaviour support it is often possible that multiple goals and actions can be recommended in a given situation. For example, in order to achieve the goal of working out, the user could either go for a run or do yoga. However, some of these options may be preferable to the user or should be prioritised to ensure the functionality of the agent.

In Default Logic, we represent this using an initial theory (K, D) and preferences over the possible outcomes. The knowledge base K of the agent describes its functionality, some of the current context and a set of plans which specify what actions must be true in order for a goal to be achieved. We can specify that the user only wants to do one workout, it is not raining, a high-intensity workout consists of going for a run or using the rowing machine and a low-intensity workout consists of yoga or going for a walk. This is the information the agent considers to be certain. The agent can reason with this to infer further knowledge, but this will usually not be enough to make a helpful recommendation. For this purpose we use the default rules, or in the following also assumption rules D to make assumptions based on previously inferred information. These rules specify that if a prerequisite φ is true and it is consistent to assume the consequence ψ, then ψ is inferred. We only allow one concept in the consequence of each rule and differentiate the rules based on this. We restrict the prerequisites of the rules to only contain context information. For example, we may have rules that state "if it is not raining and there is nothing that says otherwise, the agent will suggest going for a run" and "if there is no reason not to, the agent will suggest using the rowing machine". When introducing the knowledge base, we specified that the user only wants one workout recommendation per day. The agent can therefore not apply both assumption rules at once, even though both are currently applicable. Instead, the agent identifies multiple different scenarios, which are possible extensions of the theory. In our example there will be one extension with the rowing machine workout and one with running. In order to decide which of these actions to recommend to the user, the agent considers an ordering on these possible outcomes. These orderings can be based on the preferences of the user but also on the priority of the outcome to the functionality of the agent. In this example, the user prefers running over using the rowing machine so the agent recommends the run.

The reasoning process we propose is intended to reflect the common planning strategy that actions are selected based on the goals that must be achieved and goals are selected based on the circumstances. When computing which advice to give to the user, the agent will begin with the knowledge base K. This information is then completed using standard logic inference rules to form the theory $Th(K)$ containing all sentences that the agent is certain of. The agent then makes assumptions about the user's context by using the rules with context information in the consequences. It uses the priority ordering on the

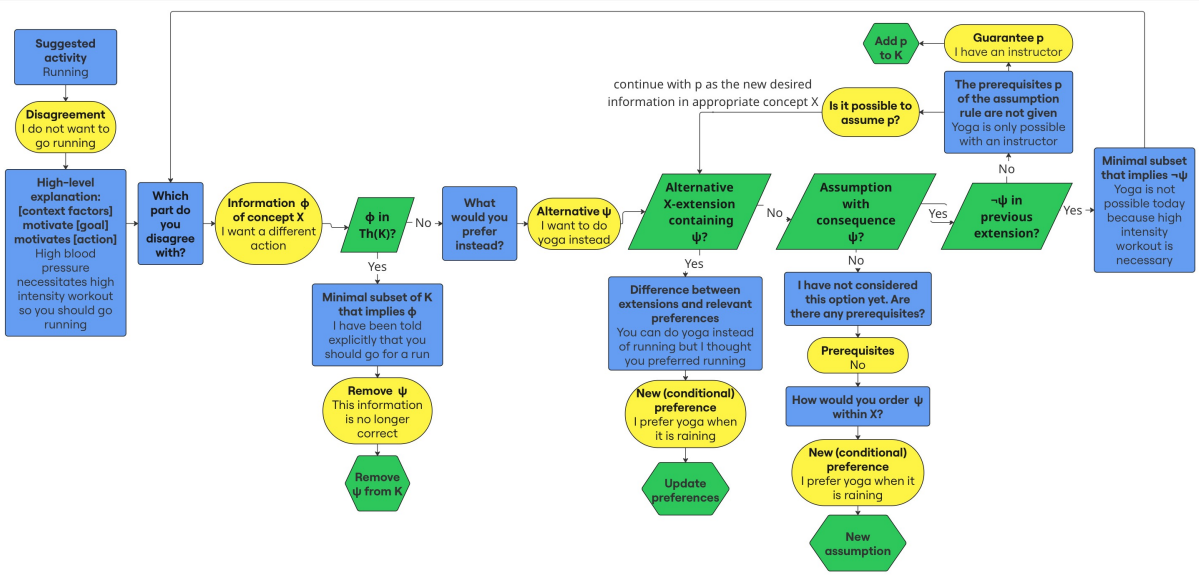

Figure 1: Flowchart of the Realignment Dialogue

outcomes to select one of these extensions. After this, the goal assumption rules are used to include additional assumptions about the goals the user should pursue and the priorities are again used to select an extension. Lastly, the action assumption rules are used to determine the possible actions that could be taken by the user while considering the previously assumptions made about the context and the goals. We note that this will always include the actions necessary in order to achieve the selected goals because we have included the plans for each goal in the knowledge base. The preference ordering on the possible actions is then used to select the final extension, which will be the basis for the agents advice. Any actions that are included in this will be recommended to the user in order to achieve the included goals.

3. Realignment Dialogues

The agent we described in the previous section is intended to give the user the desired advice but misalignments can still arise for a variety of reasons. This includes the use of inaccurate or outdated information or simply the user's preferences changing. While designing an agent it is therefore important to not only avoid misalignments altogether but to ensure there are ways the user can the agent can resolve any problems that occur while the agent is in use.

For this purpose we introduce realignment dialogues in which the user is guided through the agent's reasoning process leading to an undesired conclusion. We do not want the user to go through every step of this, so we begin by giving a general overview and then zooming in to the relevant areas until the problem is identified. During this process, the agent will also collect additional information to update its knowledge base in order to eliminate the cause of the misalignment. The outline of the interaction we propose is given in Figure 1. The blue rectangles represent the agent's explanations and questions to the user, the yellow ovals represent the input of the user, the green diamonds are internal queries the agent makes and the green hexagons represent the updates the agent makes. The bold text states the general structure, the plain text gives an example.

Each realignment dialogue begins with the agent giving a recommendation and the user disagreeing with this. The interaction ends when the agent performs an update to resolve the misalignment.

1. The agent provides a general explanation for its recommendation by stating the goal, the action this contributes to and the relevant context. We determine which context is relevant by looking at the prerequisites of the respective goal and action selection rules.
2. The user selects which piece of information φ of the concept X has been reasoned about incorrectly.

This may not necessarily be the cause of the misalignment, instead this is a slightly more detailed description of the problem.

3. The agent checks if the misaligned information φ is contained in the theory $Th(K)$.

3.1. If φ is in $Th(K)$, the agent gives the user a minimal subset of the initial knowledge base K that implies φ. The user then removes any unwanted information.

3.2. If φ is not in $Th(K)$, the agent asks the user to explain what alternative information ψ they would want the agent to consider. While this can be left as an open question, we can also offer some suggestions to help the user. If applicable, the agent can present the most preferred alternative extensions as options for the user to choose from. We also allow the user to say "anything except φ", which we interpret as wanting to enforce $\neg\varphi$. Besides this, the user can give alternate suggestions to explain their preferred outcome.

4.1. If there are any alternative possible extensions containing ψ that were not selected, the agent will present the most preferred of these alternatives to the user. In particular, the agent highlights the differences between the two suggestions and which preferences were responsible for the agent's previous decision. The user can then update these preferences.

4.2. Otherwise, the agent searches for assumption rules with ψ as the consequent. If there are none, then ψ is a new possibility for the agent. The agent will ask for the prerequisites of this new assumption rule and the ordering of ψ compared to the other outcomes of this concept. This information is then added to the knowledge base.

4.3. If there is an assumption rule with the consequence ψ, then this rule is not applicable. If the negation of ψ has been inferred in a previous step, we know this must be due to the information in the knowledge base in combination with previous assumptions. The agent gives the user a minimal subset of the information that implies $\neg\psi$. The interaction then continues with step 2. based on this explanation of the misalignment scenario.

4.4. Otherwise, the rule must be inadmissible because the prerequisites were not given. The agent tells the user the unsatisfied prerequisites of the rule in question. The user can then choose to enforce this prerequisite by adding it to the knowledge base or question whether there is a possible extension in which it is true by repeating step 4.

4. Discussion

In this paper, we have introduced an interaction between the agent and the user which is intended to identify and resolve potential misalignments. We used interactive explanations to create a shared understanding of the situation and employ the user's input to update the agent's knowledge base and resolve the misalignment.

While we have not tested the interaction in practice yet, we can begin to evaluate the explanations that are given by discussing the Co-12 properties from [14]. The correctness, completeness, consistency and continuity of our explanations is inherently given by the fact that our agent uses knowledge-based reasoning and the explanations are based on formal proofs. On the other hand, there is no probability information so we do not consider the confidence property is not applicable. Regarding contrastivity, while we do allow the user to ask questions of the form "why not x instead?", we have not included questions of the form "what if x?". This is because each change in the agent's reasoning requires us to recompute the advice that the agent will give. We have reduced the covariate complexity and coherence of our explanations by making each explanation concrete and only including context descriptors, goals and actions that the user will already be familiar with from use of the agent. In particular, we focus on the context of resolving a misalignment and only discuss the parts of the knowledge base that are involved in this. By breaking the explanation down into smaller steps within a larger conversation we increase the compactness of each individual explanation. However, it is possible that more complex examples will require additional techniques to optimise this further. In this paper we have focused on the structure of the interactive explanations rather than the individual presentation and language, so the composition and controllability of the explanations is not within the scope of this paper.

Overall, our approach to explanations seems promising regarding the Co-12 properties, while also enabling the agent to collect the information necessary to update its reasoning. Through this shared understanding of the situation, the agent and the user are able to collaborate and determine the best support. In future work we also hope to study how users experience the interaction we have designed. This includes working on the presentation of the explanations in natural language and studying whether the fixed answer options are perceived as complete.

Additionally, we have focused on the ability of the user and the agent to resolve misalignments by changing the knowledge base of the agent. However, in practice we would not want all parts of the agent's knowledge base to be changeable by the user as this could diminish the functionality of the agent or even create potential safety risks. In future work we will explore how we can protect certain unchangeable knowledge and communicate this to the user in a way that will support the acceptance of these limitations as well as the search for potential compromises. For added functionality we may also want to consider the distinction between the permanent changes to the agent's knowledge base and potential temporary exceptions which do not need to be remembered by the agent.

Acknowledgments

This research was partly funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl, grant number 024.004.022.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] H. Oinas-Kukkonen, Behavior change support systems: A research model and agenda, in: T. Ploug, P. Hasle, H. Oinas-Kukkonen (Eds.), Persuasive Technology, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 4–14. doi:10.1007/978-3-642-13226-1_3.

[2] Z. Akata, D. Balliet, M. de Rijke, F. Dignum, V. Dignum, G. Eiben, A. Fokkens, D. Grossi, K. Hindriks, H. Hoos, H. Hung, C. Jonker, C. Monz, M. Neerincx, F. Oliehoek, H. Prakken, S. Schlobach, L. van der Gaag, F. van Harmelen, H. van Hoof, B. van Riemsdijk, A. van Wynsberghe, R. Verbrugge, B. Verheij, P. Vossen, M. Welling, A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence, Computer 53 (2020) 18–28. doi:10.1109/MC.2020.2996587.

[3] M. B. van Riemsdijk, C. M. Jonker, V. Lesser, Creating socially adaptive electronic partners: Interaction, reasoning and ethical challenges, in: Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems, AAMAS '15, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2015, p. 1201–1206.

[4] P.-Y. Chen, M. Tielman, D. Heylen, C. Jonker, M. Riemsdijk, Acquiring semantic knowledge for user model updates via human-agent alignment dialogues: An exploratory focus group study, in: HHAI 2023: Augmenting Human Intellect - Proceedings of the 2nd International Conference on Hybrid Human-Artificial Intelligence, IOS Press, 2023, pp. 93–108. doi:10.3233/FAIA230077.

[5] C. Jonker, M. Riemsdijk, B. Vermeulen, Shared mental models - a conceptual analysis., in: Coordination, Organizations, Institutions, and Norms in Agent Systems VI - COIN 2010 International Workshops, 2010, pp. 132–151.

[6] J. Wolff, V. de Boer, D. Heylen, M. B. van Riemsdijk, Defining an adaptable framework for behaviour support agents in default logic, in: CEUR workshop proceedings, volume 3835, CEUR, 2024, pp. 72–82.

[7] R. Reiter, A logic for default reasoning, Artificial Intelligence 13 (1980) 81–132. URL: https://www.sciencedirect.com/science/article/pii/0004370280900144. doi:https://doi.org/10.1016/0004-3702(80)90014-4, special Issue on Non-Monotonic Logic.

[8] F. Doshi-Velez, B. Kim, Considerations for Evaluation and Generalization in Interpretable Machine Learning, Springer International Publishing, Cham, 2018, pp. 3–17. URL: https://doi.org/10.1007/978-3-319-98131-4_1. doi:10.1007/978-3-319-98131-4_1.

[9] E. H. Shortliffe, R. Davis, S. G. Axline, B. G. Buchanan, C. Green, S. N. Cohen, Computer-based consultations in clinical therapeutics: Explanation and rule acquisition capabilities of the mycin system, Computers and Biomedical Research 8 (1975) 303–320. URL: https://www.sciencedirect.com/science/article/pii/0010480975900099. doi:https://doi.org/10.1016/0010-4809(75)90009-9.

[10] B. Bogaerts, E. Gamba, T. Guns, A framework for step-wise explaining how to solve constraint satisfaction problems, Artificial Intelligence 300 (2021) 103550. URL: https://www.sciencedirect.com/science/article/pii/S0004370221001016. doi:https://doi.org/10.1016/j.artint.2021.103550.

[11] M. Harbers, K. van den Bosch, J.-J. Meyer, A study into preferred explanations of virtual agent behavior, in: Intelligent Virtual Agents, Springer, 2009, pp. 132–145.

[12] M. L. Tielman, M. C. Suárez-Figueroa, A. Jönsson, M. A. Neerincx, L. Cavalcante Siebert, Explainable ai for all - a roadmap for inclusive xai for people with cognitive disabilities, Technology in Society 79 (2024) 102685. URL: https://www.sciencedirect.com/science/article/pii/S0160791X24002331. doi:https://doi.org/10.1016/j.techsoc.2024.102685.

[13] N. Tintarev, J. Masthoff, A survey of explanations in recommender systems, in: 2007 IEEE 23rd International Conference on Data Engineering Workshop, 2007, pp. 801–810. doi:10.1109/ICDEW.2007.4401070.

[14] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, C. Seifert, From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai, ACM Computing Surveys 55 (2023). URL: https://doi.org/10.1145/3583558. doi:10.1145/3583558.