Why Bad Coffee? Explaining BDI Agent Behaviour with Valuings (Extended Abstract)[*]

Michael Winikoff [1†], **Galina Sidorenko**[2], **Virginia Dignum**[3] and **Frank Dignum**[3]
[1]Victoria University of Wellington
[2]Halmstad University
[3]Umeå University
michael.winikoff@vuw.ac.nz, galina.sidorenko@hh.se, {virginia, dignum}@cs.umu.se

Abstract

An important issue in deploying an autonomous system is how to enable human users and stakeholders to develop an appropriate level of trust in the system. It has been argued that a crucial mechanism to enable appropriate trust is the ability of a system to explain its behaviour. Obviously, such explanations need to be comprehensible to humans. Due to the perceived similarity in functioning between humans and autonomous systems, we argue that it makes sense to build on the results of extensive research in social sciences that explores how humans explain their behaviour. Using similar concepts for explanation is argued to help with comprehensibility, since the concepts are familiar. Following work in the social sciences, we propose the use of a folk-psychological model that utilises beliefs, desires, and "valuings". We propose a formal framework for constructing explanations of the behaviour of an autonomous system, present an (implemented) algorithm for giving explanations, and present evaluation results.

1 Introduction

The deployment of autonomous systems in situations where they interact with people raises the need for these systems to be able to explain their behaviour. Such explanations are important for a range of reasons. They allow humans to better understand the system in order to be able to more effectively coordinate with it and anticipate its behaviour, as well as understand its limitations.

The problem our paper [Winikoff *et al.*, 2021] therefore addresses is how an autonomous system can provide an explanation for why it chose a particular course of action. In other words, answering questions of the form "*why did you do X?*". Specifically, we develop a computational mechanism that provides building blocks for explanations that are subsequently assembled into explanations for why a particular action was performed.

Since we want explanations to be comprehensible and useful to humans, we draw on work that examines how we, as humans, explain our behaviour to each other, in particular, drawing on the seminal work of Malle [2004].

Malle argues that humans use folk psychological constructs (e.g. beliefs, desires) to explain their behaviour. There is also empirical evidence that humans use these constructs to explain the behaviour of robots [de Graaf and Malle, 2019; Thellman *et al.*, 2017]. This leads us to adopt a model that includes desires and beliefs, specifically the well-known BDI model [Rao and Georgeff, 1992; Bratman *et al.*, 1988; Bratman, 1987]. Doing so allows explanations to be based on concepts that are the same as those used by humans when explaining their actions.

Malle also provides a careful analysis, supported by experimental evidence, that highlights a number of different types of concepts used to explain behaviour. The ones that are most relevant to the context we are considering (software explaining its decisions) are what Malle terms *reasons*. Malle identifies three types of reasons: desires, beliefs, and what he terms *valuings*. He defines valuings as things that "*directly indicate the positive or negative affect toward the action or its outcome*". For example, someone appreciating snow. He goes on to argue that valuings are a third concept, distinct from either beliefs or goals. We therefore extend our BDI model with valuings, following Cranefield *et al.* [2017].

The remainder of this paper is a high-level summary of Winikoff *et al.* [2021], purposely leaving out formal definitions. We use a running example, which, as suggested by the paper's title, relates to getting coffee. Jo is an academic visiting colleagues at another University. Like many academics, he requires coffee for optimal functioning. There are a number of possible sources of coffee. The little kitchen near Ann's office has coffee-like-substance freely available, but this machine requires a staff card to operate. Ann has in her office a coffee machine which converts pods into nice coffee (we assume that Ann has a plentiful supply of coffee pods). There is also a coffee shop a few buildings away, where good coffee can be obtained, at a (financial) cost. Jo prefers coffee to coffee-like substances, which is the over-riding preference. Less-important preferences are to save money, and to use the nearest coffee source. Therefore the three relevant quality attributes are (in order): quality (coffee preferred to coffee-like), money (free preferred to expensive), and loca-

[*]Full paper in Artificial Intelligence Journal, Volume 300, 2021.
[†]Contact Author

tion (smallest distance from starting location).

2 Formal Setting

The basic formal structure we use is that of a goal tree, which indicates how the goals of an agent are achieved through subgoals and in the end by actions performed. This structure is an abstraction of a wide range of BDI agent platforms where agents are specified using plans which have a trigger, a context condition, and a plan body.

A *goal tree* is a tree of nodes, where leaves (nodes with no children) are actions (each action having associated pre- and post-conditions), and non-leaf nodes can be decomposed using either AND, SEQ, or OR. The children of an OR decomposed node are *options* and each have an associated condition that indicates in what situation the option can be used.

Intuitively, a goal tree is executed as follows. If the tree is simply an action, then the action is performed (assuming its pre-conditions hold), resulting in its post-condition. If the tree is an AND or SEQ decomposition, then all of the sub-goals are executed, either in the specified sequential order (SEQ), or in some, unspecified, order (AND). Finally, if the tree is an OR decomposition, then an applicable option (i.e. one whose condition is currently believed to hold) is selected and executed.

We now turn to valuings. We incorporate them by specifying a preference over options[1], which can depend on key aspects of their effects. For instance, we prefer high-quality coffee, so annotating the different options with the quality of coffee that is obtained allows the preference for higher-quality coffee to be expressed.

Figure 1 shows a goal tree for obtaining coffee. Following the running example, it shows three alternative ways of getting coffee: getting bad coffee from the kitchen (getKitchenCoffee), getting good coffee from Ann's office (getOfficeCoffee), and getting very good, but expensive, coffee from the shop (getShopCoffee). Each of these options have children, in this case (mostly) actions that are done in sequence. We also specify (not shown in the figure) that we prefer getOfficeCoffee and getShopCoffee to getKitchenCoffee (since getKitchenCoffee provides only a coffee-like substance). However, the preference between getOfficeCoffee and getShopCoffee depends on the context (e.g. how much money one has, and the distance of each from the current location).

3 Generating Explanations

A central contribution of the AIJ paper is a definition of the different explanatory factors and a function E that takes a question ("why did you do A?") and, given a goal tree G and an execution trace T, yields a set of explanatory factors.

We now briefly define the structure of an explanation, and then go on to informally present the explanation function E.

An explanation is given in terms of reasons which can be desires (goals), beliefs, or valuings (expressed as "I preferred V to $\{V_1, \ldots, V_n\}$"). In addition to these reasons, which follow Malle, We also have forward-looking explanatory factors of the form "I did N_1 in order to be able to later do N_2", and factors of the form "I tried N but it failed".

Our presentation of the explanation function E proceeds in a number of steps. We first explain how we identify belief and preference explanatory factors. We then add preparatory factors, motivations (desires), and dealing with failure handling. For space reasons we do not discuss filtering of the explanatory factors (see [Winikoff *et al.*, 2021, §3.6]).

Intuitively, we define the set of belief and valuing explanatory factors for a goal-tree using the following cases, where the question being answered is "why did you do A when executing goal-tree G?" and the trace of actions that were done is T. If A was not done (i.e. $A \notin T$) then the explanation is just $\perp$ (a special form denoting that the question does not make sense). Otherwise the following cases apply.

- If the root of G is A then the explanation is the preconditions of A. The intuition is that in order for A to have been done, its pre-conditions had to hold, so they are part of the explanation for why A was done.
- Otherwise, we include the explanation of those children of G that are the roots of sub-trees that have some nodes that appear in the trace T prior to A (recursively calling E with each child node). The intuition is that if a child of G does not have any sub-node that appears in the trace prior to A, then it could not have influenced whether A was done, so it does not need to be considered.
- Additionally, if G is OR decomposed, then we also include an additional explanation relating to why the particular option taken was chosen. This additional explanation has three parts, and is complex, reflecting that the choice between options is at the heart of the endeavour of explanation.
    1. The condition of the option that was selected. Intuitively, in order for the option to be selected, its condition had to be true, so is part of the explanation.
    2. The conditions of each option that could *not* have been selected.
    3. For each condition that was not selected but that could have been selected, an indication that the selected sub-goal was preferred to it in the current situation.

To illustrate, consider the scenario given in §1 and assume that Ann is not in her office (i.e. C_2 is false), but C_1 and C_3 are true. Then the preference explanation is (in English): "I chose to get coffee from the shop because [1.] I had money, and [2.] Ann was not in her office, and [3.] I prefer V_3 to V_1 in this situation".

On the other hand, in a situation where all C_i are true (so any option can be selected) and C_3 is selected, the explanation would take the form (in English): "I chose to get coffee from the shop because [1.] I had money, and [3.] I prefer V_3 to both V_1 and V_2 in this situation" (since all options could be selected the second case "[2.]" is empty and is elided).

We now extend the definition to also include preparatory actions. For example, an explanation for "why did you go to the kitchen?" could also be "because I need to be in the

[1] The subtle distinctions between "value", "valuings", and "preferences" and their relationship are discussed in the AIJ paper [Winikoff *et al.*, 2021, §2].

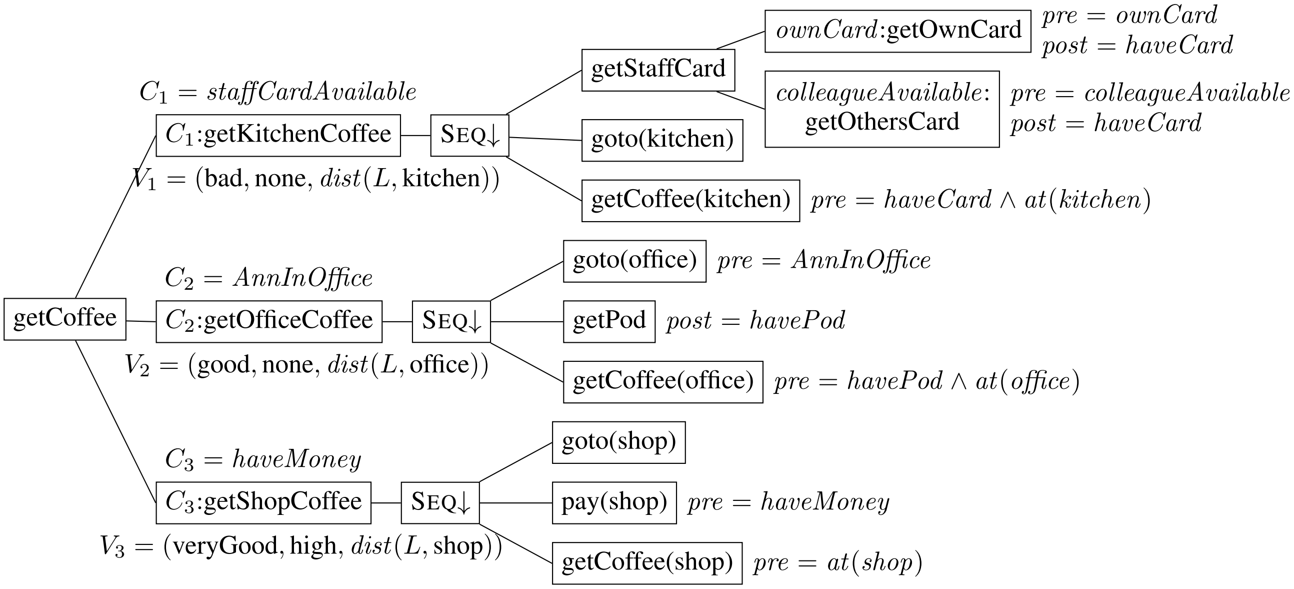

Figure 1: Running Example. Notation: option nodes are written $C{:}N$ where C is the condition and N the node name; V_i are value effect annotations of the form (quality, cost, distance), where $dist(L_1, L_2)$ is the distance between locations L_1 and L_2; children are OR-refined, except where indicated with a SEQ, in which case they are ordered from top to bottom; finally, for any X the post-condition of $goto(X)$ is $at(X)$ and the post-condition of $getCoffee(X)$ is $haveCoffee$.

kitchen in order to get coffee". This is where an action's post condition is (a necessary part of) the pre-condition of a future action. Specifically, a preparatory reason applies to explain an action A_1 when (i) the post-condition of A_1 is required in order for the pre-condition of another action A_2 to hold, and (ii) A_2 only occurs after A_1 (to be precise: if A_2 occurs, then it must be preceded by A_1). In order to define this we define that there is a *link* between A_1 and A_2 if these two conditions hold, and generalise this by considering transitive links, and links between goals. In order to define links between goals we define pre- and post-conditions of goals by aggregating the pre- and post-conditions of their children.

We next extend the explanatory function E to also include explanatory factors that explain the agent's motivations (i.e. desires). For instance, in the running example one factor explaining why an agent went to the shop would be "... and I desired to getShopCoffee". We extend the explanatory function E with these motivations by including particular ancestors[2] of the node A that is the subject of the question "why did you do A?".

Finally, we extend E to include explanatory factors relating to failure handling. Informally, actions can fail, and the failure of a node is handled by considering its parent (this mechanism is common to many BDI languages). If the parent is a SEQ or an AND then it too is considered to be failed, and failure handling moves to consider that node's parent. When an OR node is reached, failure is handled by trying an alternative plan (if one exists, otherwise the OR node is deemed to have failed).

In explanation, our key idea for dealing with failed nodes is to exclude them from the generation of explanatory factors outlined so far, and instead, we add a fourth case to the explanation of an OR decomposed node: an indication of the options that were tried but failed ("... and I already unsuccessfully tried doing X"). To illustrate this consider a situation where Jo has decided to getOfficeCoffee, but by the time he reaches Ann's office, Ann has had to leave for a meeting. The plan therefore fails, and Jo then recovers by electing to go to the shop. In response to the query "why did you getShopCoffee?" the explanation given is (in English): "because I have money, I prefer good coffee to bad coffee, and because I tried (and failed) to get pod coffee".

4 Implementation

The AIJ paper goes on to define an efficient ($O(N)$) algorithm to implement the explanation function E. The paper also provides an algorithm for rendering explanations in English. In addition to mapping the explanatory factors to English text, this algorithm also deals with generating a specific explanation for why a given option was preferred to another, taking into account the various factors and their importance.

For example, if we know that one of the factors is the most important, e.g. coffee quality, and we preferred *getShopCoffee* to *getKitchenCoffee*, then we can say: "The made choice (getShopCoffee) has the best quality of coffee and that is the most important attribute". On the other hand, if we know that the most important attribute can be overridden by other factors, then for example when we preferred *getKitchenCoffee* to *getShopCoffee*, we can say: "I chose getKitchenCoffee with poorer quality of coffee despite quality of coffee being my

[2] We include all ancestors of A that are *not* OR refined. The reason is that for an OR refined goal, the parent goal is redundant, since the child goals convey more information: they capture the specific approach taken to achieve the parent goal.

most important attribute, because it was cheaper and closer than the other option (getShopCoffee)".

We now give an example to illustrate the English generation done by the two algorithms. Consider a scenario where the selected option is to go to the shop, and where the starting location was close to the shop (so there was a low distance to the shop, and a higher distance to the office and the kitchen). Given this scenario, let us first assume that there is a single most-important attribute, namely the quality of the coffee. In this case the answer to "why did you pay(shop)?" is:

> "I had Money; *getShopCoffee has the best quality of coffee (and that is the most important attribute)*; I needed to pay shop in order to getCoffee shop; I desired to getShopCoffee"

On the other hand, if we assume that instead the quality and price are both equally important (but that distance is less important), then the italic sentence above would instead be: "getShopCoffee is preferred because it has better quality and distance than getOfficeCoffee and getKitchenCoffee"

5 Evaluation

In order to assess the comprehensibility and usability of the explanations generated, as well as provide guidance to future work on selecting explanations, we conducted two human participant evaluations. For the full details on these studies see Winikoff *et al.* [2018] for the first study, and Winikoff and Sidorenko [2021] for the second study.

Our first evaluation took the coffee scenario described and administered a survey. Participants, who were recruited on mechanical Turk, were provided with a brief description of the scenario and an indication of what behaviour was observed (participants were randomly allocated into one of three groups, each of which was given a different observed behaviour). Each of the 109 participants was given five possible explanations for the observed behaviour, and was asked to rank the explanations from most to least preferred. The first explanation (E1) combined valuings and beliefs, and corresponds to the initial explanatory function described earlier, without any extensions (e.g. when the second option is selected "*This is the best possible coffee available; I had no money*"). The second and third explanations are solely in terms of valuings. The second (E2) is abstract, just saying literally "*This is the best possible coffee available*", while the third (E3) is concrete, with a specific explanation (e.g. "*This coffee is better than the kitchen and cheaper than in the shop*"). The fourth candidate explanation (E4) provides only relevant beliefs (e.g. "*I've no money; Ann was in her room*"). Finally, the fifth candidate explanation (E5) gives the goal, and the beliefs that enabled the specific behaviour that was selected, which is the explanation mechanism proposed by Harbers [2011] (e.g. "*I wanted coffee; Ann was in her room*").

We found that explanations E1 and E3 were considered better than the other explanations (all p values < 0.05), and that, except for Group 2, E2 was seen as being the worst[3]. Since E1 and E3 both include valuings, this indicates that valuings are of value in providing effective explanations. Furthermore, for Groups 2 and 3, E1 was preferred to E3, indicating that valuings alone were not sufficient.

The second evaluation also used a survey. Participants were recruited using advertisements in a range of undergraduate lectures within the Otago Business school, by email to students at co-authors' institutions, and by posting on social media. The scenario used involved a software personal assistant. Each participant was presented with five possible explanations given in a random order, and was asked to rank the explanations from most to least preferred. The explanations combine different elements of the explanation mechanism described earlier. Specifically, there are four types of elements that can be included in an explanation: beliefs, valuings, desires, and links. Explanation E1 includes all four elements, explanation E2 includes only valuings and beliefs, E3 includes only valuings, E4 includes only beliefs, and E5 includes only beliefs and desires.

We found that respondents preferred explanations which have valuing, belief, and desire components. They did not prefer explanations that have links. Furthermore, the analysis showed that of the four factors, the presence of valuing components most strongly (and significantly) correlates with higher preference for the explanation. In other words, explanations including valuings are more likely to be preferred. These results are consistent with, and reinforce, the findings from the first evaluation.

6 Conclusion

We have argued that explaining the behaviour of autonomous software could be done using the same concepts as are used by humans when explaining their behaviour. We then proposed a formal framework, using BDI-style goal-trees, augmented with value effect annotations. This formal framework was used to define an explanation function, which has been implemented and evaluated. That our explanations are couched in terms of familiar concepts suggests that the explanations are likely to be comprehensible. This is seen in the presented example explanations, which are not excessively long, nor excessively complex. More importantly, this claim is also supported by the evaluation results, including the finding in both evaluations that valuings are seen as important components of good explanations. There are a range of directions for future work, and we refer the reader to the closing section of the AIJ paper [Winikoff *et al.*, 2021, §7] for these.

Acknowledgements

We would like to thank Dr Damien Mather, at the University of Otago, for statistical advice relating to the second evaluation. This work was supported by a University of Otago Research Grant (UORG), and some of it was done while the first two authors were at the University of Otago. We would also like to thank the anonymous reviewers of the AIJ submission for their detailed and insightful comments that helped to improve the paper. Finally, we were amused, after the publication of the paper, to be invited to contribute a chapter to "*The Global Aspects of Coffee Production - Value Chains and Contribution to Sustainable Development*".

[3] For Group 2 the behaviour observed was getting coffee from the shop, which meant that E2 was in fact a good explanation.

References

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

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

[Cranefield *et al.*, 2017] Stephen Cranefield, Michael Winikoff, Virginia Dignum, and Frank Dignum. 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*, pages 178–184, 2017.

[de Graaf and Malle, 2019] Maartje M. A. de Graaf and Bertram F. Malle. People's Explanations of Robot Behavior Subtly Reveal Mental State Inferences. In *14th ACM/IEEE International Conference on Human-Robot Interaction, HRI 2019, Daegu, South Korea, March 11-14, 2019*, pages 239–248. IEEE, 2019.

[Harbers, 2011] Maaike Harbers. *Explaining Agent Behavior in Virtual Training*. SIKS dissertation series no. 2011-35, SIKS (Dutch Research School for Information and Knowledge Systems), 2011.

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

[Rao and Georgeff, 1992] Anand S. Rao and Michael P. Georgeff. An abstract architecture for rational agents. In C. Rich, W. Swartout, and B. Nebel, editors, *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning*, pages 439–449, San Mateo, CA, 1992. Morgan Kaufmann Publishers.

[Thellman *et al.*, 2017] Sam Thellman, Annika Silvervarg, and Tom Ziemke. Folk-Psychological Interpretation of Human vs. Humanoid Robot Behavior: Exploring the Intentional Stance toward Robots. *Frontiers in Psychology*, 8:1–14, 2017.

[Winikoff and Sidorenko, 2021] Michael Winikoff and Galina Sidorenko. Evaluating a mechanism for explaining BDI agent behaviour. https://profwinikoff.files.wordpress.com/2021/07/evaluating.pdf, 2021. Accessed: 2022-06-01.

[Winikoff *et al.*, 2018] Michael Winikoff, Virginia Dignum, and Frank Dignum. Why bad coffee? Explaining agent plans with valuings. In Barbara Gallina, Amund Skavhaug, Erwin Schoitsch, and Friedemann Bitsch, editors, *Computer Safety, Reliability, and Security - SAFECOMP 2018 Workshops, ASSURE, DECSoS, SASSUR, STRIVE, and WAISE, Västerås, Sweden, September 18, 2018, Proceedings*, volume 11094 of *Lecture Notes in Computer Science*, pages 521–534. Springer, 2018.

[Winikoff *et al.*, 2021] Michael Winikoff, Galina Sidorenko, Virginia Dignum, and Frank Dignum. Why bad coffee? explaining BDI agent behaviour with valuings. *Artif. Intell.*, 300:103554, 2021.