Evaluating a Mechanism for Explaining BDI Agent Behaviour

Michael Winikoff[1(✉)] and Galina Sidorenko[2]

[1] Victoria University of Wellington, Wellington, New Zealand
michael.winikoff@vuw.ac.nz
[2] Halmstad University, Halmstad, Sweden
galina.sidorenko@hh.se

Abstract. Explainability of autonomous systems is important to supporting the development of appropriate levels of trust in the system, as well as supporting system predictability. Previous work has proposed an explanation mechanism for Belief-Desire-Intention (BDI) agents that uses folk psychological concepts, specifically beliefs, desires, and valuings. In this paper we evaluate this mechanism by conducting a survey. We consider a number of explanations, and assess to what extent they are considered believable, acceptable, and comprehensible, and which explanations are preferred. 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, whereas those explanations that include links are least likely to be preferred. We also found evidence that single-factor explanations, as used in some previous work, are too short.

Keywords: Explainable Agency · Belief-Desire-Intention (BDI) · Evaluation

1 Introduction

> "*Explainability is crucial for building and maintaining users' trust in AI systems*." [16]
> "*Automated systems should provide explanations that are technically valid, meaningful and useful to you and to any operators or others who need to understand the system*" https://www.whitehouse.gov/ostp/ai-bill-of-rights/, published 4 October 2022.

It is now widely accepted that explainability is crucial for supporting an appropriate level of trust in autonomous and intelligent systems (e.g. [12,16,29]). However, explainability is not just important to support (appropriate) trust. It also makes a system understandable [34], which in turn allows systems to be challenged, to be predictable, to be verified, and to be traceable [34].

In this paper we focus on *autonomous agents*: software systems that are able to act autonomously. This includes a wide range of physically embodied systems

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

D. Calvaresi et al. (Eds.): EXTRAAMAS 2023, LNAI 14127, pp. 18–37, 2023.
https://doi.org/10.1007/978-3-031-40878-6_2

(e.g. robots) and systems that do not have physical embodiment (e.g. smart personal assistants) [25,26,28]. Although autonomous systems use AI techniques, not all AI systems are autonomous, e.g. a system may be simply making recommendations to a human, rather than taking action itself.

Explainability is particularly important for autonomous systems [20,36], since, by definition, they take action, so, depending on the possible consequences of their actions, there is a need to be able to trust these systems appropriately, and to understand how they operate. One report proposes to include "*...for users of care or domestic robots a why-did-you-do-that button which, when pressed, causes the robot to explain the action it just took*" [32, Page 20]. It has also been argued that explainability plays an important role in making autonomous agents accountable [8].

However, despite the importance of explainability of autonomous systems, most of the work on explainable AI (XAI) has focused on explaining machine learning (termed "data-driven XAI" by Anjomshoae *et al.* [4]), with only a much smaller body of work focusing on explaining autonomous agents (termed "goal-driven XAI" by Anjomshoae *et al.* [4], and "explainable agency" by Langley *et al.* [20]). Specifically, a 2019 survey [4] found only 62 distinct published papers on goal-driven XAI published in the period 2008–2018.

In order to develop a mechanism for an autonomous agent to be able to answer in a useful and comprehensible way questions such as "*why did you do X?*", it is useful to consider the social sciences [23]. In particular, we draw on the extensive (and empirically-grounded) work of Malle [21]. Malle argues that humans use folk psychological constructs in explaining their behaviour[1]. Specifically, in explaining their behaviour, humans use the concepts of beliefs, desires[2], and valuings[3].

Prior work [38] has used these ideas to develop a mechanism that allows Belief-Desire-Intention (BDI) agents [5,6,27] (augmented with a representation for valuings, following [9]) to provide explanations of their actions in terms of these concepts.

In this paper we conduct an empirical human subject evaluation of this mechanism, including an evaluation of the different component types of explanations (e.g. beliefs, desires, valuings). Such evaluations are important in assessing the effectiveness of explanatory mechanisms. For example, are explanations using beliefs seen as less or more preferred than explanations that use desires, or that use valuings? Empirical evaluation can answer these questions, and by answering them, guide the development and deployment of explanation mechanisms for autonomous agents. Specifically, the key research question we address[4] is: ***What forms of explanation of autonomous agents are preferred?***.

[1] There is also empirical evidence that humans use these constructs to explain the behaviour of robots [13,33].

[2] Terminology: we use "goal" and "desire" interchangeably.

[3] Defined by Malle as things that "*directly indicate the positive or negative affect toward the action or its outcome*". Whereas values are generic (e.g. benevolence, security [30]), valuings are about a given action or outcome. Valuings can arise from values, but can also be directly specified without needing to be linked to higher-level values. In our work we represent valuings as preferences over options [38, Sect. 2].

[4] We also consider (Sect. 4.4) the question: "to what extent is trust in a given system determined by a person's more general attitudes towards technology, and towards Artificial Intelligence?".

An earlier evaluation of this explanation mechanism has been conducted [37] (the results of which are also briefly summarised in [38]). However, this paper differs from the earlier evaluation in that: (i) we use a different scenario, (ii) we use different patterns of explanations, including links (which were not included in the earlier evaluation), (iii) we also include questions on trust in technology, and (iv) we conduct a deeper and more sophisticated analysis, including an assessment of the effects of the different explanatory component types, and of the correlation between trust in the autonomous system and more general trust in technology.

We propose a number of hypotheses, motivated by existing literature (briefly indicated below, and discussed in greater length in Sect. 5). Our hypotheses all relate to the *form* of the explanation. Since the explanation we generate has four types of explanatory factors, we consider for each of these types how they are viewed by the user (H1–H3). Furthermore, since including more types of explanatory factors results in longer explanations, we also consider the overall effect of explanation length (H4).

H1: Explanations that include valuings are more likely to be preferred by users over other forms of explanations (that do not include valuings). This hypothesis is based on the finding of [37].

H2: Explanations that include desires are more likely to be preferred by users over explanations that include beliefs. This hypothesis is based on the findings of [7, 15, 17] (discussed in detail in Sect. 5).

H3: Explanations that include links are *less* likely to be preferred by users over other forms of explanations (that do not include links). This hypothesis is based on the findings of [15].

H4: Shorter explanations are more likely to be preferred by users. This hypothesis is based on the arguments of (e.g.) [17]. Note that they argued that explanations ought to be short, and therefore only evaluated short explanations. In other words, their evaluation did not provide empirical evidence for this claim.

The remainder of this paper is organised as follows. We begin by briefly reviewing the explanation mechanism that we evaluate (Sect. 2). Next, Sect. 3 presents our methodology, and then Sect. 4 presents our results. We finish with a review of related work (Sect. 5), followed by a brief discussion (Sect. 6) summarising our findings, noting some limitations, and indicating directions for future work.

2 Explanation Mechanism

We now briefly review the explanation mechanism. For full details, we refer the reader to [38]. In particular, here we focus on the *form* of the explanations, omitting discussion of *how* the explanations are generated.

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: (i) Walking, (ii) Cycling, if a bicycle is available, and (iii) Catching a bus, if money is available (i.e. there is enough credit on your card). One particular afternoon, you are about to leave to go home, when the*

E1: A bicycle was not available, money was available, 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, I needed to buy a bus ticket in order to allow you to go by bus, and I have the goal to allow you to catch the bus.
E2: A bicycle was not available, money was available, and 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.
E3: 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.
E4: A bicycle was not available, and money was available.
E5: A bicycle was not available, money was available, and I have the goal to allow you to catch the bus.

Fig. 1. Explanations E1–E5

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.

An explanation is built out of four types of building blocks: desires, beliefs, valuings, and links.

- A **desire (D)** explanation states that the agent having a certain desire was part of the reason for taking a certain action. For example, that the system chose to buy a bus ticket because it desired to allow you to catch the bus.
- A **belief (B)** explanation states that the agent having a certain belief was part of the reason for taking a certain action. For example, that the system chose to buy a ticket because it believed that a bicycle was not available.
- A **valuing (V)** explanation states that the agent chose a certain option (over other options) because it was *valued*. For example, that the system chose to select catching a bus because it was the fastest of the available options, and that getting home more quickly is valued.
- Finally, a **link (L)** explanation states that a particular action was performed in order to allow a subsequent action to be done. For example, that the agent bought the ticket in order to allow the user to then catch the bus (which requires having a ticket).

A full explanation may use a number of each of these elements, 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 have the goal to allow you to catch the bus (D).*

3 Methodology

We surveyed[5] participants[6], who were recruited using advertisements in a range of undergraduate lectures within the Otago Business school, by email to students at institutions of two colleagues, Frank and Virginia Dignum, with whom we were collaborating on related work, and by posting on social media. New Zealand based participants were given the incentive of being entered into a draw for a NZ$100 supermarket voucher.

The scenario used the software personal assistant ("SAM") explained in Sect. 2.

Each participant is presented with five possible explanations (see Fig. 1) which are given in a random order, i.e. each participant sees a different ordering. The explanations combine different elements of the explanation mechanism described earlier in this paper. 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 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[7] 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 also asked to indicate whether they would like further clarification of the explanation given, for instance, by entering into a dialog with the system, or providing source code.

Once all five explanations were considered, participants were asked to rank the explanations from 1 (most preferred) to 5 (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.* [22, 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.

4 Results

We received 74 completed responses to the online survey. The demographic features of the respondents are shown in Table 1.

[5] The survey can be found at: https://www.dropbox.com/s/ec6fg3u1rqhytcb/Trust-Autonomous-Survey.pdf.

[6] Ethics approval was given by University of Otago (Category B, D18/231).

[7] Where 1 was labelled "Strongly Disagree", 7 was labelled "Strongly Agree", and 2–6 were not labelled.

Table 1. Selected demographic characteristics of respondents (percentage distributions; percentages may not sum to 100% due to rounding)

Characteristic		Percentage
Gender	male	55.4
	female	41.9
	not answered	2.7
Age	18–24	39.2
	25–34	27.0
	35–44	14.9
	45–54	14.9
	55–64	4.0
Education	High school graduate	17.6
	Bachelor/undergraduate degree	44.6
	PhD degree/Doctorate	36.5
	not answered	1.4
Ethnicity	New Zealander (non Māori)	31.1
	Māori	2.7
	European	46.0
	Other	20.3

4.1 Analysis of Believability, Acceptability and Comprehensibility of Explanations

We begin by analysing how participants assessed each of the explanations E1-E5 on three characteristics: Believability, Acceptability and Comprehensibility. Each explanation was assessed on its own (in random order), i.e. the participants in this part of the survey were not asked to compare explanations, but to assess each explanation in turn.

The descriptive statistics regarding the Believability, Acceptability and Comprehensibility of the five Explanations are shown below (recall that 1 is "strongly disagree" and 7 is "strongly agree", so a higher score is better).

We used paired Wilcoxon-signed rank tests to estimate differences in means. The results are given in Table 2. These results show that most of the differences between pairs of explanations in terms of their Believability, Acceptability, and Comprehensibility are statistically significant[8] with $p < 0.005$.

Characteristic	Explanation	Mean	Std. Dev.	Median
Believability	E1	3.90	1.78	4
	E2	4.80	1.50	5
	E3	5.08	1.34	5
	E4	3.73	1.87	4
	E5	3.76	1.72	4
Acceptability	E1	5.12	1.70	5
	E2	5.14	1.52	5
	E3	4.57	1.74	5
	E4	3.76	1.95	4
	E5	4.45	1.81	5
Comprehensibility	E1	5.55	1.38	6
	E2	5.77	1.03	6
	E3	5.62	1.04	6
	E4	4.99	1.63	5
	E5	4.85	1.64	5

Figure 2 depicts the relationships in Table 2. For believability (top left of Fig. 2) explanations E3 and E2 are statistically significantly different to explanations E1, E4 and E5 (in fact E3 and E2 are better than E1, E4 and E5 since they have a higher median). However, E3 and E2 are not statistically significantly different to each other, nor are there statistically significant differences amongst E1, E4 or E5. For acceptability (bottom of Fig. 2) the situation is a little more complex: explanations E1 and E2 are statistically significantly different to the other three explanations[9] (but not to each other), and E3 and E5 are both statistically significantly better than E4 (but E3 and E5 are not statistically significantly different). Finally, for comprehensibility (top right of Fig. 2), explanations E2, E3 and E1 are statistically significantly different to explanations E4 and E5, but for each of the two groups of explanations there are not statistically significant differences within the group.

Overall, considering the three criteria of believability, comprehensibility, and acceptability, these results indicate that E2 is statistically significantly better than E4 and E5 according to all criteria, and is statistically significantly better than E1 (Believability only), and E3 (Acceptability only). Explanation E3 was statistically significantly better than E4 (all criteria), E5 (Believability and Comprehensibility), and

[8] We use 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.

[9] Although for E1-E3 it is only at $p = 0.0273$.

Table 2. Statistical Significance of Differences in means for Believability, Acceptability and Comprehensibility. Bold text indicates statistical significance with $p < 0.005$ and "***" indicates $p < 0.0001$.

Characteristic	Explanation	E1	E2	E3	E4	E5
Believability	E1	–	***	***	0.6006	0.6833
	E2	***	–	0.2015	***	***
	E3	***	0.2015	–	***	***
	E4	0.6006	***	***	–	0.9808
	E5	0.6833	***	***	0.9808	–
Acceptability	E1	–	0.7357	0.0273	***	***
	E2	0.7357	–	**0.0041**	***	***
	E3	0.0273	**0.0041**	–	**0.0003**	0.6481
	E4	***	***	**0.0003**	–	**0.0002**
	E5	***	***	0.6481	**0.0002**	–
Comprehensibility	E1	–	0.1275	0.7370	**0.0040**	**0.0022**
	E2	0.1275	–	0.1510	***	***
	E3	0.7370	0.1510	–	**0.0005**	**0.0005**
	E4	**0.0040**	***	**0.0005**	–	0.6060
	E5	**0.0022**	***	**0.0005**	0.6060	–

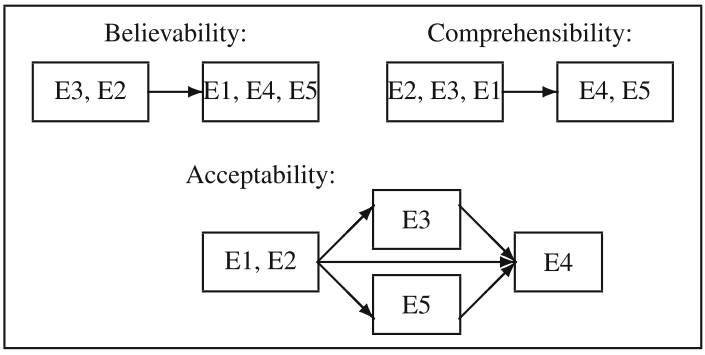

Fig. 2. Visual representation of the significance results in Table 2 where an arrow indicates a statistically significant difference (arrow is directional from better to worse)

E1 (Believability). Explanation E1 was statistically significantly better than E4 and E5 (Comprehensibility and Acceptability), and E3 (Acceptability). Finally, E5 is better than E4 (Acceptability only).

So, 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).

4.2 Analysis of Rankings of Explanations

The analysis below relates to the part of the survey where respondents were asked to rank a set of five explanations from 1 (most preferred) to 5 (least preferred).

To analyse the ranked data we employed a general discrete choice model (linear mixed model), using a ranked-ordered logit model which is also known as an exploded logit [3].

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.

The ranked-ordered logit is based on a multistage approach where the standard logit [3] is applied to the most preferred choice J_1 in the set of all alternatives $(J_1, \ldots, J_K)$, then to the second-ranked choice J_2 in the set $(J_2, \ldots, J_K)$ after the first-ranked item was removed from the initial choice set and so on.

The ranked-ordered logit model was estimated with the SAS procedure PHREG, yielding results shown below. Each row (e.g. row E2) is in relation to the reference explanation, E1. The column β gives the key parameter, showing the relative likelihood. These estimates indicate that, on average, respondents are most likely to prefer explanation E2 ($\beta_{E2} = 0.475$) and least likely to prefer E4 ($\beta_{E4} = -1.077$). The odds of preferring E2 are $exp^{0.475} = 1.608$ times the odds of preferring E1. The right-most column ("Pr > ChiSq") shows that the β value for each explanation except for E3 is statistically significantly different to that of E1.

Explanation	β	Standard Error	Chi-Square	Pr > ChiSq
E2	0.475	0.166	8.18	0.0042
E3	−0.154	0.165	0.878	0.3488
E4	−1.077	0.17	40.016	<.001
E5	−0.887	0.168	28.034	<.0001

We also calculated the Wald chi-square for all the possible pairs or coefficients (see below). All but two of the tests were statistically significant[10], with p-values less than 0.005 (actually less than 0.001). The two non-significant pairs were E1–E3 and E4–E5, which were not significant at the 0.005 level.

[10] As before, we use a significance level of 0.005 rather than 0.05 to avoid type II errors, given the number of tests performed.

Label	Wald Chi-Square	Pr > ChiSq
$\beta_{E2} - \beta_{E3}$	14.1768	0.0002
$\beta_{E2} - \beta_{E4}$	77.3522	<.0001
$\beta_{E2} - \beta_{E5}$	61.7307	<.0001
$\beta_{E3} - \beta_{E4}$	28.8808	<.0001
$\beta_{E3} - \beta_{E5}$	18.9785	<.0001
$\beta_{E4} - \beta_{E5}$	1.3091	0.2526
$\beta_{E2} - \beta_{E1}$	8.1801	0.0042
$\beta_{E3} - \beta_{E1}$	0.8780	0.3488
$\beta_{E4} - \beta_{E1}$	40.0157	<.0001
$\beta_{E5} - \beta_{E1}$	28.0341	<.0001

This analysis therefore allows us to conclude that, based on participants ranking of the explanations, 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 of the previous section.

In order to provide additional confidence in the logit analysis, we also performed a series of comparisons between pairs of items using a Wilcoxon signed rank test. This also found that all differences were significant at the 0.005 level, except for the two pairs that were not significantly different at this level according to the regression analysis. Thus, the exploded logit model gives results that are qualitatively the same as those obtained by a standard nonparametric method.

We also investigated whether there are differences between males and females in their ranking of explanations. Using the same exploded logit model and new dummy variable for gender, we computed the Wald chi-square statistic for the null hypothesis that differences between gender-dependent coefficients are zero, which had p-value 0.95. Thus, there is no evidence for a difference between men and women in ranking explanations. A similar analysis was made for age-dependent groups of respondents

Table 3. The construction of the explanations.

Component	E1	E2	E3	E4	E5
B(eliefs)	1	1	0	1	1
V(aluings)	1	1	1	0	0
D(esires)	1	0	0	0	1
L(inks)	1	0	0	0	0
Length in words:	63	36	27	9	20
Length in characters:	318	206	152	54	101

and found no significant difference in ranking of explanations in relation to age (p-value 0.158).

4.3 Effects of Explanation Components

Next, we investigated the effects of explanation components (e.g. beliefs, desires, valuings) and how they affect ranking. There were four possible components: beliefs, valuings, desires and links. The constructed explanations are shown in Table 3 where ones indicate the presence of respective components and zeros indicate their absence. For example, the first column indicates that explanation E1 has all four components, whereas the second column shows that E2 has only the beliefs and valuings components.

As shown in Table 4, all except one of the coefficients of the exploded logit model are significantly different from zero at level $p = 0.005$ and the only exception β_D, corresponding to desires, is significant at the 0.05 level. A positive coefficient indicates that this component is more preferred, whereas a negative coefficient indicates that the component is less preferred. Thus, respondents prefer explanations that have V, B, and D components. They are reluctant to prefer explanations that have links. The magnitudes of coefficients in Table 4 can be interpreted as follows. The presence of V components in the explanation has produced $100 \times (\exp^{\beta} - 1) = 100 \times (\exp^{2.4} - 1) = 1002.3$ percent increase in the odds of preferring this explanation to the one where V is absent, controlling for other components. The presence of beliefs in the explanation has produced $100 \times (\exp^{0.82} - 1) = 127$ percent increase in the odds of preferring this explanation to the one where B is absent, controlling for other components. The presence of desires in the explanation has produced $100 \times (\exp^{0.54} - 1) = 71.6$ percent increase in the odds of preferring this explanation to the one where D is absent, controlling for other components. For links we have $100 \times (\exp^{-1.16} - 1) = -68.65\%$, which implies that the odds of preferring explanation with links over the one where L is absent goes down by 68.65%.

Table 4. Respondents' Preferences in Ranking Components V,B,D,L: Analysis of Maximum Likelihood Estimates

Parameter	Parameter Estimate (β)	Standard Error	Chi-Square	Pr > ChiSq
V	2.402	0.224	115.28	<.0001
B	0.821	0.176	21.661	0.0001
D	0.543	0.224	5.88	0.0153
L	−1.164	0.285	16.6224	0.0001

As before, we also calculated the Wald chi-square for all the possible pairs or coefficients. We found that the difference between preferring B and D is not statistically significant ($p = 0.33$), whereas the difference among all others components is significant (see Table 5).

Table 5. Statistically Significant Differences in regression coefficients

Label	Wald Chi-Square	Pr > ChiSq
$\beta_V - \beta_B$	52.2652	<.0001
$\beta_V - \beta_D$	90.3121	<.0001
$\beta_V - \beta_L$	56.0711	<.0001
$\beta_B - \beta_D$	0.9473	0.3304
$\beta_B - \beta_L$	27.2910	<.0001
$\beta_D - \beta_L$	12.5446	0.0004

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. In other words, explanations including valuings are more likely to be preferred.

4.4 Analysis of Overall Trust in SAM

Our final analysis considered the relationship between overall trust in a specific autonomous system (SAM), and broader trust in technology in general, and AI specifically. The question being addressed here is: to what extent is trust in a given system, such as SAM, determined by a person's more general attitudes towards technology, and towards Artificial Intelligence?

As noted earlier, the survey included 11 questions that assessed three dimensions of attitudes [22]: faith in technology (4 questions), general attitude to technology (3 questions), and attitude to Artificial Intelligence (4 questions).

We conducted a reliability analysis to assess the internal consistency of these blocks of questions. The results (see Table 6) show that the Cronbach's alpha coefficients ranged from[11] 0.73 to 0.85. We also considered all of the questions taken together ("Merged" in Table 6), which yielded a higher alpha. This meant that the questions forming the components of the scale were sufficiently intercorrelated to allow the dimensions to be merged. We therefore merged the three dimensions into a single item that measured each participant's attitude to technology in general (including AI).

In order to assess the extent to which broader background attitudes to technology influenced trust in SAM we compared the calculated background trust measure (average of the ten questions) against each participant's response to the question "I trust SAM because it can provide me a relevant explanation for its actions" (Likert response on a 1–7 scale).

To estimate the correlation between background trust in technology and trust in SAM, we calculated Spearman's coefficient. The coefficient value of 0.46 confirms that

[11] For the AI group of questions, the analysis indicated that dropping the third question would improve the alpha from 0.69 to 0.79, which was done, meaning that we used a total of 10 questions. The dropped question was: "I think that current problems with use of AI (bias, breach of privacy, etc.) will be solved in the short term".

Table 6. Analysis of dimensions of background trust to technology

Characteristic	Cronbach's alpha
Faith	0.73
General attitude to technology	0.85
Attitude to AI	0.79
Merged	0.91

there appears to be a positive correlation between the two variables ($\rho_S = 0.46, n = 74$, $p = 3.8 \times 10^{-5}$). Thus, high values of background trust in technology are associated with high "trust in SAM" scores.

Interestingly, although the correlation is clearly significant ($p = 3.85 \times 10^{-5}$), it is not that strong ($\rho_S = 0.46$, which is considered a moderate strength correlation). In other words, knowing that a person has, say, a high level of trust in technology in general, does not allow one to confidently predict that they will therefore have a high level of trust in an autonomous system (see Fig. 3). In other words, trust in autonomous systems is not purely determined by background trust in technology more broadly.

We also assessed the effects of gender. A Wilcoxon test performed for two independent groups (men and women) showed no evidence for a difference in means for SAM score (W = 551.5, p-value = 0.33). So, we can conclude that there is no evidence that men and women give different scores to SAM.

5 Related Work

As noted in the introduction, there is comparatively little work on goal-driven XAI. Focusing specifically on approaches that use beliefs and desires, and that conduct an evaluation, there are a number of papers.

Harbers *et al.* [7,14,15] consider an explanation mechanism that is similar to the one we evaluate in that it uses explanation templates that correspond to our explanatory components of beliefs, desires, and links. However, they do not have a corresponding template for valuings. Furthermore, their explanations do not take into account possible alternatives, i.e. they explain why X was done solely in terms of what *enabled* X to be done, rather than considering why X was *selected* from amongst the available options. In general, X may be enabled, but whether it is selected can depend also on the availability of other options. For example, choosing to catch a bus because a bicycle is not available, so cycling (which otherwise would be preferred) is not an option. An explanation in terms of what enabled us to catch a bus (having money), is not useful. A useful explanation in this scenario is that the preferred option (cycling) was not available due to the lack of a bicycle being available.

Turning to the evaluations, Broekens *et al.* [7] report on an evaluation using a cooking domain. They had 30 participants who were randomly allocated to one of the three explanation types. Participants were asked to score an explanation for each action in

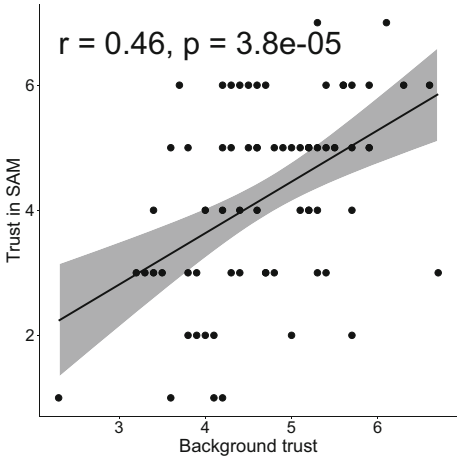

Fig. 3. Correlation between "trust in SAM" score and background trust.

terms of naturalness[12] and usefulness[13]. They found that, in general, goal-based explanations were preferred. However, the specific preferred explanation depended on the action and its context. For example, where an action is an "or" (i.e. its parent goal requires a single child to be selected), then a belief-based explanation is more helpful.

Harbers *et al.* [15] report on an evalation using a fire-fighting domain, with 20 participants who were not experts in the domain. For each action, they asked participants which of four explanations was preferred: the parent goal (in the goal hierarchy tree[14]), the parent's parent goal, the beliefs, and a link explanation. Similarly to Broekens *et al.* they found that the choice depended on the action and its context. However, in general, links were barely selected as preferred, and while goals were well-received, for "or" actions beliefs were preferred.

These results are consistent with ours in that we also found that links were not preferred. One difference is that while their explanations consisted of a single type (e.g. belief or goal or link), we considered more complex explanations that mixed elements. And, of course, they did not consider valuings, so our key finding, that valuings are more preferred than either belief-based or goal-based explanations, was not able to be identified by their work.

Kaptein *et al.* [17,19] considered explanations in the context of an e-health application. In earlier work [17] they evaluated user preferences for explanations in the context of a personal assistant that worked with a fictitious child ("jimmy") who has type 1 diabetes mellitus. Participants (19 adults and 19 children) were provided with a number of scenarios, and asked to select their preferred explanation for each one. The expla-

[12] Explained as: "With a natural explanation we mean an explanation that sounds normal and is understandable, an explanation that you or other people could give.".

[13] Explained as: "Indicate how useful the explanations would be for you in learning how to make pancakes.".

[14] The tree of goals, beliefs, and actions.

nations given as options were either a single belief, or a single goal. In both cases the explanation provided was the belief/goal immediately above the action in the goal hierarchy tree. This ensured that the explanation was short (a single element). They found that both children and adults preferred goal-based explanations, and that adults had a stronger preference for these than children. However, they caution that the preference between goals and beliefs can depend on context, and in particular, that in their work the participants were already considered to be familiar with the domain, since the children participating in the evaluation themselves had type 1 diabetes.

In later work [19] Kaptein *et al.* evaluated whether the form of the explanation provided affected the *behaviour* of children with type 1 diabetes using an e-health support system. A distinguishing feature of this evaluation is that it was conducted "in the wild" over a longer time period (2.5–3 months), with 48 children[15] aged 6–14. As in the previous evaluation, explanations were kept short, being either a single belief or single goal ("cognitive" explanations), or an emotional explanation ("affective" explanations). The emotional explanations were obtained by rephrasing from e.g. "I want to ..." to "It would make me happy if you ...". They found only a single statistically significant result, which was counter-intuitive: providing explanations (either cognitive or affective) correlated with children following the tasks *less* often. The authors hypothesised a number of possible explanations for this behaviour, for example, that children read the explanation, and if the aim of the task is to teach them something that they already believe they know, then they are therefore less likely to select that task.

Again, these results are consistent with ours, in that we found varying preference between beliefs and desires. However, as noted for Harbers *et al.*, their explanations did not mix explanation types, and they did not consider valuings. On the other hand, they included affective explanations, which were not part of our evaluation.

More recently, Abdulrahman *et al.* [1,2] conducted an empirical human subject study to assess explanations provided by an intelligent virtual advisor. Their study was limited to university students (mostly under 20 years old), with 91 participants. It concerned a virtual assistant ("Sarah") that was designed to give advice to help students manage stress. Like us, they drew inspiration from Malle, but they did not include valuings in their explanations. They considered explanations that contained beliefs only, desires only, and both beliefs and desires[16]. The key question they consider is to what extent "*... do explanations that refer to the user's beliefs or goals influence the user's intention to change the behaviours recommended by the agent?*". They did not find a difference between belief-only and goal-only explanations, but found that belief-and-goal explanations did not lead to a significant change in intentions to join a study group (the recommendation from the agent), which they ascribe to the explanation being longer.

Mualla *et al.* [24] propose an explanation mechanism focussed on parsimony, which requires balancing brevity and adequacy of the explanation. They use contrastive explanations and different forms of filtering to attempt to provide parsimonious explanations. Their evaluation, which is done using a scenario involving understanding UAV operations, hypothesises that using contrastive rather than only normal explanations, and

[15] One child was excluded from the data analysis due to a data glitch.

[16] Since their virtual assistant was only providing advice, rather than performing a sequence of actions, it did not make sense to have link explanations.

adaptive rather than static filtering, both improve understandability of explanations. They divided participants into three groups: normal explanations and static filtering (SF), normal explanations and adaptive filtering (AF), and adaptive filtering with both normal and contrastive explanations (AC). Comparing survey results for these groups they found that while adaptive filtering on its own was not necessarily better (AF vs. SF), the combination of adaptive filtering and contrastive explanation did make a significant difference (SF vs. AC). They also evaluated trust, but did not find any statistically significant relationship regarding the effect of explanation type on trust. This last point can perhaps be explained by our finding that trust is to some extent influenced by background trust in technology: if the effect of explanations on trust is only partial (since trust is also influenced by other factors, such as trust in technology), then we might expect to see that the effect on trust of changing the *form* of the explanation would not be statistically significant. Our findings regarding the length of explanations support their argument for parsimony: our most preferred explanation was neither the longest nor the shortest. Finally, we note that their explanation mechanism does not include valuings, and that our results suggest that it should.

6 Discussion

We have conducted a human participant empirical evaluation of explanations of BDI agents, where the explanations consist of different types of explanatory components: beliefs, desires, valuings, and links.

We found that participants assess the different explanations somewhat differently for Believability, Acceptability, and Comprehensibility, and that most of the differences between the assessment of different explanations were statistically significant (Sect. 4.1). Overall, considering both assessing each explanation on its own (Sect. 4.1) and explicitly ranking the explanations (Sect. 4.2), we have a consistent preference for E2 (which has belief and valuing explanatory components), followed by E1 (all component types) and E3 (valuing only), which are not distinguishable from each other. The least preferred explanations were E4 (belief only) and E5 (belief and desire), which are also not distinguishable from each other in terms of preferences.

Analysing the data to assess preferences for the different types of explanatory components (beliefs, desires, valuings, links; see Sect. 4.3), we found that the presence of valuing components make an explanation significantly more likely to be preferred, and that the presence of belief and/or desire components also makes an explanation more likely to be preferred, but less so than valuings. On the other hand, the presence of a link component makes an explanation less likely to be preferred.

Finally (Sect. 4.4), there is statistically significant correlation between trust in SAM and trust in technology in general ($p = 3.85 \times 10^{-5}$), but the correlation has moderate strength ($\rho_S = 0.46$). 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.

Returning to our hypotheses, we have that:

H1: Explanations that include valuings are more likely to be preferred by users over other forms of explanations (that do not include valuings). This hypothesis is confirmed by our findings (Sect. 4.1, 4.2 & 4.3).

H2: Explanations that include desires are more likely to be preferred by users over explanations that include beliefs. This hypothesis is **not** confirmed: we did not find a statistically significant difference between preferences for beliefs and desires (Sect. 4.3).

H3: Explanations that include links are less likely to be preferred by users over other forms of explanations (that do not include links). This hypothesis is confirmed by our findings (Sect. 4.3).

H4: Shorter explanations are more likely to be preferred by users. Interestingly, this hypothesis is **not** confirmed: explanations E1 (the longest, with all four types of explanatory factors) and E3 (with only a single factor) did not have a statistically significant difference in preference (Sect. 4.2). Indeed, E1 was considered more acceptable than E3, whereas E3 was considered more believable than E1 (Sect. 4.1). Furthermore, there was not a significant difference in their comprehensibility (Sect. 4.1). Indeed, the two least-preferred explanations (E4 and E5) were the shortest!

Based on these findings, we provide the following advice to guide the development of explanations.

Firstly, it is clear that valuings are valued. Explanations that included a valuing component (E1, E2 and E3) were significantly more likely to be preferred. This is consistent with the findings of the previous evaluation [37], which also found that valuings were valued[17]. We therefore recommend that when developing explanation mechanisms based on this framework, that valuing explanatory factors are included in explanations.

Secondly, we found that explanations including link components were less likely to be preferred. The evaluation by Harbers *et al.* [15] also found that link explanations were barely selected as preferred. However, we exercise a note of caution: 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 (Sect. 4.2) 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 to limit explanations to a single belief or a single goal. We therefore recommend that when providing explanations, the explanations are not limited to only single factors. Furthermore, when evaluating forms

[17] Specifically, their explanations corresponding in structure with our E2 (valuing and belief) and E3 (valuing) were most preferred.

of explanation, longer explanations should also be considered and included in the evaluation.

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.* [18] argue that explanations should include emotions. This is an interesting idea, and one that would be good to investigate further. It would also be good to consider other evaluation metrics such as relevance and the extent to which explanations relate to what the user already knows. Finally, our evaluation only considered explanations that were presented to the user all at once. It would also be good to consider explanations that are 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. [10,11,31,35]).

Acknowledgements. 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. Abdulrahman, A., Richards, D., Bilgin, A.A.: Reason explanation for encouraging behaviour change intention. In: Dignum, F., Lomuscio, A., Endriss, U., Nowé, A. (eds.) AAMAS 2021: 20th International Conference on Autonomous Agents and Multiagent Systems, Virtual Event, United Kingdom, 3–7 May 2021, pp. 68–77. ACM (2021). https://www.ifaamas.org/Proceedings/aamas2021/pdfs/p68.pdf
2. Abdulrahman, A., Richards, D., Bilgin, A.A.: Exploring the influence of a user-specific explainable virtual advisor on health behaviour change intentions. Auton. Agents Multi Agent Syst. **36**(1), 25 (2022). https://doi.org/10.1007/s10458-022-09553-x
3. Allison, P.D., Christakis, N.A.: Logit models for sets of ranked items. Sociol. Methodol. **24**, 199–228 (1994). https://www.jstor.org/stable/270983
4. Anjomshoae, S., Najjar, A., Calvaresi, D., Främling, K.: Explainable agents and robots: results from a systematic literature review. In: Elkind, E., Veloso, M., Agmon, N., Taylor, M.E. (eds.) Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2019, Montreal, QC, Canada, 13–17 May 2019, pp. 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems (2019). https://dl.acm.org/citation.cfm?id=3331806
5. Bratman, M.E., Israel, D.J., Pollack, M.E.: Plans and resource-bounded practical reasoning. Comput. Intell. **4**, 349–355 (1988)
6. Bratman, M.E.: Intentions, Plans, and Practical Reason. Harvard University Press, Cambridge (1987)
7. Broekens, J., Harbers, M., Hindriks, K., van den Bosch, K., Jonker, C., Meyer, J.-J.: Do you get it? user-evaluated explainable BDI agents. In: Dix, J., Witteveen, C. (eds.) MATES 2010. LNCS (LNAI), vol. 6251, pp. 28–39. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-16178-0_5
8. Cranefield, S., Oren, N., Vasconcelos, W.W.: Accountability for practical reasoning agents. In: Lujak, M. (ed.) AT 2018. LNCS (LNAI), vol. 11327, pp. 33–48. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-17294-7_3
9. Cranefield, S., Winikoff, M., Dignum, V., Dignum, F.: 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, pp. 178–184 (2017). DOI: https://doi.org/10.24963/ijcai.2017/26

10. Dennis, L.A., Oren, N.: Explaining BDI agent behaviour through dialogue. In: Dignum, F., Lomuscio, A., Endriss, U., Nowé, A. (eds.) AAMAS 2021: 20th International Conference on Autonomous Agents and Multiagent Systems, Virtual Event, United Kingdom, 3–7 May 2021, pp. 429–437. ACM (2021), https://www.ifaamas.org/Proceedings/aamas2021/pdfs/p429.pdf
11. Dennis, L.A., Oren, N.: Explaining BDI agent behaviour through dialogue. Auton. Agents Multi Agent Syst. **36**(1), 29 (2022). https://doi.org/10.1007/s10458-022-09556-8
12. Floridi, L., et al.: Ai4people–an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds Mach. **28**(4), 689–707 (2018). https://doi.org/10.1007/s11023-018-9482-5
13. de Graaf, M.M.A., Malle, B.F.: 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, 11–14 March 2019, pp. 239–248. IEEE (2019). https://doi.org/10.1109/HRI.2019.8673308
14. Harbers, M.: Explaining agent behavior in virtual training. SIKS dissertation series no. 2011–35, SIKS (Dutch Research School for Information and Knowledge Systems) (2011)
15. Harbers, M., van den Bosch, K., Meyer, J.C.: Design and evaluation of explainable BDI agents. In: Huang, J.X., Ghorbani, A.A., Hacid, M., Yamaguchi, T. (eds.) Proceedings of the 2010 IEEE/WIC/ACM International Conference on Intelligent Agent Technology, IAT 2010, Toronto, Canada, 31 August–3 September 2010, pp. 125–132. IEEE Computer Society Press (2010). https://doi.org/10.1109/WI-IAT.2010.115
16. High-Level Expert Group on Artificial Intelligence: The assessment list for trustworthy artificial intelligence (2020). https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment
17. Kaptein, F., Broekens, J., Hindriks, K.V., Neerincx, M.A.: 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, 28 August–1 September 2017, pp. 676–682. IEEE (2017). https://doi.org/10.1109/ROMAN.2017.8172376
18. Kaptein, F., Broekens, J., Hindriks, K.V., Neerincx, M.A.: 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, 23–26 October 2017, pp. 88–93. IEEE Computer Society (2017). https://doi.org/10.1109/ACIIW.2017.8272595
19. Kaptein, F., Broekens, J., Hindriks, K.V., Neerincx, M.A.: 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, 3–6 September 2019, pp. 1–7. IEEE (2019). https://doi.org/10.1109/ACII.2019.8925526
20. Langley, P., Meadows, B., Sridharan, M., Choi, D.: Explainable agency for intelligent autonomous systems. In: Singh, S., Markovitch, S. (eds.) Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, California, USA, 4–9 February 2017, pp. 4762–4764. AAAI Press (2017). https://aaai.org/ocs/index.php/IAAI/IAAI17/paper/view/15046
21. Malle, B.F.: How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction. The MIT Press, Cambridge (2004). ISBN 0-262-13445-4
22. Mcknight, D.H., Carter, M., Thatcher, J.B., Clay, P.F.: Trust in a specific technology: an investigation of its components and measures. ACM Trans. Manag. Inf. Syst. **2**(2), 12:1–12:25 (2011). https://doi.org/10.1145/1985347.1985353
23. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. **267**, 1–38 (2019). https://doi.org/10.1145/1824760.1824761

24. Mualla, Y., et al.: The quest of parsimonious XAI: a human-agent architecture for explanation formulation. Artif. Intell. **302**, 103573 (2022). https://doi.org/10.1016/j.artint.2021.103573
25. Müller, J.P., Fischer, K.: Application impact of multi-agent systems and technologies: a survey. In: Shehory, O., Sturm, A. (eds.) Agent-Oriented Software Engineering, pp. 27–53. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-642-54432-3_3
26. Munroe, S., Miller, T., Belecheanu, R., Pechoucek, M., McBurney, P., Luck, M.: Crossing the agent technology chasm: experiences and challenges in commercial applications of agents. Knowl. Eng. Rev. **21**(4), 345–392 (2006)
27. Rao, A.S., Georgeff, M.P.: An abstract architecture for rational agents. In: Rich, C., Swartout, W., Nebel, B. (eds.) Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning, pp. 439–449. Morgan Kaufmann Publishers, San Mateo (1992)
28. van Riemsdijk, M.B., Jonker, C.M., Lesser, V.R.: Creating socially adaptive electronic partners: Interaction, reasoning and ethical challenges. In: Weiss, G., Yolum, P., Bordini, R.H., Elkind, E. (eds.) Conference on Autonomous Agents and Multiagent Systems (AAMAS), pp. 1201–1206. ACM (2015). https://dl.acm.org/citation.cfm?id=2773303
29. Robinette, P., Li, W., Allen, R., Howard, A.M., Wagner, A.R.: Overtrust of robots in emergency evacuation scenarios. In: Bartneck, C., Nagai, Y., Paiva, A., Sabanovic, S. (eds.) The Eleventh ACM/IEEE International Conference on Human Robot Interation, HRI 2016, Christchurch, New Zealand, 7–10 March 2016, pp. 101–108. IEEE/ACM (2016). https://doi.org/10.1109/HRI.2016.7451740
30. Schwartz, S.: An overview of the Schwartz theory of basic values. Online Read. Psychol. Cult. **2**(1), 11 (2012). https://doi.org/10.9707/2307-0919.1116
31. Sklar, E.I., Azhar, M.Q.: Explanation through argumentation. In: Imai, M., Norman, T., Sklar, E., Komatsu, T. (eds.) Proceedings of the 6th International Conference on Human-Agent Interaction, HAI 2018, Southampton, United Kingdom, 15–18 December 2018, pp. 277–285. ACM (2018). https://doi.org/10.1145/3284432.3284470
32. The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems: Ethically Aligned Design: A Vision For Prioritizing Wellbeing With Artificial Intelligence And Autonomous Systems, Version 1. IEEE (2016). https://standards.ieee.org/develop/indconn/ec/autonomous_systems.html
33. Thellman, S., Silvervarg, A., Ziemke, T.: Folk-psychological interpretation of human vs. humanoid robot behavior: exploring the intentional stance toward robots. Front. Psychol. **8**, 1–14 (2017). https://doi.org/10.3389/fpsyg.2017.01962
34. Verhagen, R.S., Neerincx, M.A., Tielman, M.L.: A two-dimensional explanation framework to classify AI as incomprehensible, interpretable, or understandable. In: Calvaresi, D., Najjar, A., Winikoff, M., Främling, K. (eds.) EXTRAAMAS 2021. LNCS (LNAI), vol. 12688, pp. 119–138. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-82017-6_8
35. Winikoff, M.: Debugging agent programs with "Why?" questions. In: Proceedings of the 16th Conference on Autonomous Agents and Multiagent Systems, pp. 251–259 (2017)
36. Winikoff, M.: Towards trusting autonomous systems. In: El Fallah-Seghrouchni, A., Ricci, A., Son, T.C. (eds.) EMAS 2017. LNCS (LNAI), vol. 10738, pp. 3–20. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91899-0_1
37. Winikoff, M., Dignum, V., Dignum, F.: Why bad coffee? explaining agent plans with valuings. In: Gallina, B., Skavhaug, A., Schoitsch, E., Bitsch, F. (eds.) SAFECOMP 2018. LNCS, vol. 11094, pp. 521–534. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99229-7_47
38. Winikoff, M., Sidorenko, G., Dignum, V., Dignum, F.: Why bad coffee? explaining BDI agent behaviour with valuings. Artif. Intell. **300**, 103554 (2021). https://doi.org/10.1016/j.artint.2021.103554