Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)
Special Track on Human-Centred AI

The Role of Perception, Acceptance, and Cognition in the Usefulness of Robot Explanations

Hana Kopecka[1] , **Jose Such**[1,2] , **Michael Luck**[3]
[1]King's College London
[2]VRAIN, Universitat Politècnica de València
[3]University of Sussex
{hana.kopecka, jose.such}@kcl.ac.uk, michael.luck@sussex.ac.uk,

Abstract

It is known that when interacting with explainable autonomous systems, user characteristics are important in determining the most appropriate explanation, but understanding which user characteristics are most relevant to consider is not simple. This paper explores such characteristics and analyses how they affect the perceived usefulness of four types of explanations based on the robot's mental states. These types are belief, goal, hybrid (goal and belief) and baseline explanations. In this study, the explanations were evaluated in the context of a domestic service robot. The user characteristics considered are the perception of the robot's rationality and autonomy, the acceptance of the robot and the user's cognitive tendencies. We found differences in perceived usefulness between explanation types based on user characteristics, with hybrid explanations being the most useful.

1 Introduction

Explainability is of great importance in AI systems, particularly for the evaluation of fairness [Ferrer *et al.*, 2021], user trust [Mohseni *et al.*, 2018], transparency [van Nuenen *et al.*, 2020], privacy [Such, 2017], user empowerment [Abdul *et al.*, 2018] and effective control [Nunes and Jannach, 2017]. Although there are factors influencing *what* constitutes a good explanation and how to communicate it [Ribera and Lapedriza, 2019; Miller, 2019; Sanneman and Shah, 2022; Robbemond *et al.*, 2022], explanations for end-users should fit into the conceptual framework people use to explain human behaviour [De Graaf and Malle, 2017]. Some argue for the use of the folk psychological concepts of beliefs and desires [De Graaf and Malle, 2017], especially as people use these concepts themselves to explain AI systems explanations [De Graaf and Malle, 2019]. Beliefs represent one's knowledge of reality, while desires capture the preferred outcome or goal of the action [Malle, 2011] and, while they are both *reasons* for action that can be cited as explanations, they have their unique properties [Malle, 2011], as we detail later. In fact, these concepts have already inspired agent architectures like the well-known BDI agent architecture [Rao and Georgeff, 1995].

Previous work studied AI explanations based on goals or beliefs and some human factors influencing user preferences for them, e.g., [Kaptein *et al.*, 2017; Kopecka *et al.*, 2024; Harbers *et al.*, 2010a]. While some studies specifically focused on the role of user characteristics for the preference for goal or belief, in particular being an adult or a child [Kaptein *et al.*, 2017], gender, religious and political affiliation, education and cognitive style [Kopecka *et al.*, 2024], it has been hypothesised that robots should combine goals and beliefs in explanations [Harbers *et al.*, 2010a; Kopecka *et al.*, 2024; Kaptein *et al.*, 2017], as both seem to provide useful information [Harbers *et al.*, 2010a; Kaptein *et al.*, 2017]. Although recent work has included hybrid explanations in empirical studies [Winikoff and Sidorenko, 2023], they focused on further enriching them with more information. However, as far as we know, no previous work has systematically compared hybrid against only either belief or goal explanations, nor has it considered the effect of several personal characteristics on explanation preference.

In this paper, we explore the perceived usefulness of four types of explanations: (1) the *belief* that prompted the action; (2) the *goal* being pursued; (3) a *hybrid* explanation consisting of both the belief and goal; and (4) a *baseline* explanation merely re-stating the action. We aim to assess whether users prefer hybrid explanations and if user characteristics are associated with those preferences. In particular, we investigate the perceived usefulness of these types of explanations in the context of domestic service robots and their relationship with several user characteristics, such as cognitive factors (need for cognition, perception of causality and locus of attention), their acceptance of a robot (attitude towards the robot and robot anxiety) and their perception of the robot's rationality and autonomy.

To investigate the preference for explanation types and the possible differences according to cognitive tendencies or robot acceptance and perception, we formulate the following research questions, which are examined using a quantitative online survey with 468 participants:

1. What is the overall usefulness of the four types of tested explanations (belief, goal, hybrid and baseline explanation)?
2. Is acceptance and perception of the robot associated with explanation preference?

3. Are cognitive factors, such as perception of causality, locus of attention and need for cognition, associated with explanation preference?

By answering them, we provide the following contributions. We identify whether there are any differences between the perceived usefulness of the different explanation types overall and whether there are differences between a user's rating of explanation usefulness based on their cognitive tendencies and their acceptance and perception of the domestic service robot. In particular, we found that hybrid explanations are the most useful explanation type in general. Additionally, some factors, such as attitude towards the robot, were found to have a similar effect across explanation types, which is associated with finding all explanation types more useful. Other factors, such as perception of causality, are connected with changes in some explanation types.

2 Background & Related Work

2.1 Cognitive Factors and Acceptance in XAI

Cognitive factors. Cognitive tendencies can affect users' interaction with AI/robots and explanations. For example, the need for cognition, which is the tendency to enjoy effortful thinking [Cacioppo and Petty, 1982], affects how often and in which situations users require an explanation when using a music recommenders [Millecamp *et al.*, 2020]. People with a low need for cognition also benefit the most from recommendation explanations, because explanations increase their confidence while explaining the recommendation to users with a high need for cognition could lead to a decrease in confidence [Millecamp *et al.*, 2019]. However, users with low need for cognition were found to pay less attention to explanations provided by an intelligent tutoring system [Conati *et al.*, 2021]. Other cognitive characteristics were also found to influence human-AI interactions, such as attitude towards risk, computer self-efficacy, motivations, information processing style, and learning style [Anderson *et al.*, 2021].

Robot acceptance. There is no literature focused on the relationship between robot acceptance and the perception of robot explanations, but there is evidence that robot acceptance influences other aspects of human-robot interaction. For instance, in the context of retail service robots, a positive attitude towards the robot predicted higher anticipated service quality [Song and Kim, 2022], while robot anxiety was found to predict users' tendencies to avoid interacting with the robot by talking to them [Nomura *et al.*, 2008; De Graaf and Ben Allouch, 2013]. A high need for cognition has also been found to be a strong predictor of positive attitude towards service robots [Reich and Eyssel, 2013].

2.2 Goals & Beliefs in Human/AI Explanations

In human explanations. According to the folk psychology of intentional action, people make use of mental states, such as beliefs, desires and intentions, when perceiving or explaining intentional behaviour [Malle and Knobe, 1997]. Based on beliefs and desires, people create intentions, which in turn bring about their actions and, as such, beliefs and desires are understood as the reasons for intentional action, but they have different conceptual, psychological and strategic properties [Malle, 2011]. Desires represent one's desired outcome of an action and as such they represent what the person wants. Desires are relatively easy to infer from the actions themselves, as desires are often constrained by contexts and cultural scripts, which means that people usually have similar goals in common contexts. In contrast, beliefs are more difficult to infer from actions, and culture puts fewer constraints on beliefs in comparison to desires. Beliefs are formed by one's deliberate consideration of the relevant factors in the environment, desirable outcomes, possible actions and their causal relations. Belief reasons imply more deliberation than mere wanting, which is expressed by desires. Citing belief reasons might thus be motivated by the desire to portray an agent (human or robot) as a rational entity [Malle, 2011].

In AI explanations. The folk psychological concepts of beliefs, desires and intentions have been relevant to the AI and robotics communities in two important ways: as an architecture for robot reasoning and as concepts to be used in explaining robot's behaviour. The BDI (Belief-Desire-Intention) agent architecture, which is inspired by folk psychological practical reasoning, also operates with the concept of belief, which represents the knowledge of the agent about its environment, desires, the objectives of the agent and the intention, which is the course of action the agent is committed to [Rao and Georgeff, 1995]. Since BDI agents are built according to practical reasoning in humans, they can generate explanations using beliefs and desires [Harbers *et al.*, 2010b; Kaptein *et al.*, 2017]. Indeed, it has been shown that people use these concepts also to explain the behaviour of robots, suggesting that robots should explain themselves using goals and beliefs [De Graaf and Malle, 2019]. To be consistent with the terminology used in AI, we use *goal* instead of *desire* for the remainder of this paper, as desires are often conceptualised as goals in AI.

3 Hypotheses

Following best practice in empirical research [Nosek *et al.*, 2018], we registered our hypotheses with the Open Science Framework (OSF) before data collection. This is crucial to avoid mistaking testing of predictions with generation of postdictions [Nosek *et al.*, 2018] and for transparency and reproducibility of the entire research process. The registration can be accessed at https://osf.io/4smfj/?view_only=235c302c99c4492a9d0ecf2f89bf2ed2[1].

3.1 Types of Explanation in General

First, we test the overall usefulness of the explanation types. Informed by [Kopecka *et al.*, 2024; Harbers *et al.*, 2010a], we hypothesise goals and beliefs to perform similarly, hybrid explanations to be the most useful as they are the most informative and combine the strength of the two explanation types, and baseline explanation to be the least useful since they do not offer any additional information.

- H 1.1 Overall, there is no difference in preference between belief and goal explanations. Baseline explana-

[1]Note that this paper is a part of a wider cross-cultural project and only some of the hypotheses regarding the UK sample, as detailed later, are relevant for this paper.

tions are the least preferred type of explanation and hybrid explanations are the most preferred type of explanation.

3.2 Robot Acceptance and Perception

One reason people choose to cite a belief reason over a goal reason is to manage impression – explaining an action of another person citing a belief portrays them as more rational [Malle, 2011], because beliefs often represent the perceived circumstances, considered alternatives [Malle, 1999], and they are the results of an agent's deliberation over their knowledge of the environment, their desired outcomes and the causal relationship between them [Malle, 2011]. Following from this, we expect people not only to use beliefs strategically to create an impression of rationality but conversely also to prefer to receive belief explanations if they view a robot to be a rational entity, as the 'rational' belief reason would resonate with their perception of the robot and perhaps a goal explanation might thus seem inadequate. We expect the same to be true for perception of a robot's autonomy, as some believe that rationality enables and enhances autonomy [Pugh, 2020], and so the perception of rationality and autonomy might also be related.

- H 2.1 High perceived rationality and autonomy of the robot are associated with belief explanation preference.

In a similar vein, we expect to find that those who have negative attitudes and high anxiety about a robot to prefer goal explanations, which enables them to monitor the behaviour of the robot and ensure it does not have any unacceptable and harmful goals, as adopting a goal implies that the agent (robot or human) endorses a particular outcome [Malle, 2011]. This might be preferred to contextual information (belief) when one feels negative and anxious about the robot.

- H 2.2 High acceptance (positive attitude and low anxiety) towards a robot is associated with belief explanation preference.

3.3 Cognitive Factors

Another set of factors investigated for their association with explanation preference are cognitive factors describing tendencies in perceiving and processing information. As described above, goals are easier to infer because they are constrained by social scripts and often revealed by an action itself, while beliefs contain more specific information that is harder to infer [Malle, 2011] and as such is less readily available. The following hypothesis suggests that those who enjoy effortful cognitive activity (high need for cognition) prefer belief explanations as beliefs are less obvious than goals.

- H 3.1 Higher need for cognition is associated with a preference for belief explanations.

Other cognitive tendencies investigated in this paper are the perception of causality and locus of attention, which are subscales of cognitive style [Choi *et al.*, 2007]. Cognitive styles are broad tendencies describing how one thinks; these tendencies are often referred to as analytic and holistic cognitive styles. People with an analytic cognitive style focus on prominent objects, which they easily discern from the environment. Analytic thinkers notice the attributes and dispositions of these objects [Varnum *et al.*, 2010] and they attribute causality to them (dispositionism) [Varnum *et al.*, 2010; Choi *et al.*, 2007]. In contrast, holistic thinkers attend to the entire field and focus on the relationship between the prominent objects and actors and their environment [Varnum *et al.*, 2010]. People with holistic cognitive style assign causality to the interaction between the actors and the situational factors (interactionism)[Varnum *et al.*, 2010; Masuda and Nisbett, 2001; Choi *et al.*, 2007].

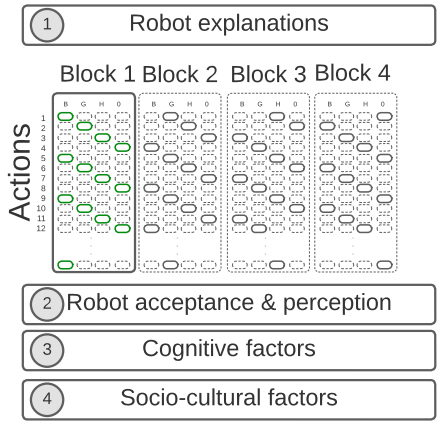


Figure 1: Schematic representation of the questionnaire.

On one hand, attention to the broader field and the interaction between actor and their environment resonates with beliefs, while focusing on salient actors and their motivation has conceptual touchpoints with goal reasons. On the other hand, Kopecka et al. [2024] report on the evidence that holistic thinkers prefer goal explanations and analytic thinkers prefer belief explanations, thus contradicting the relation hypothesised above. Due to this ambiguity, the following hypothesis does not propose any particular direction of association.

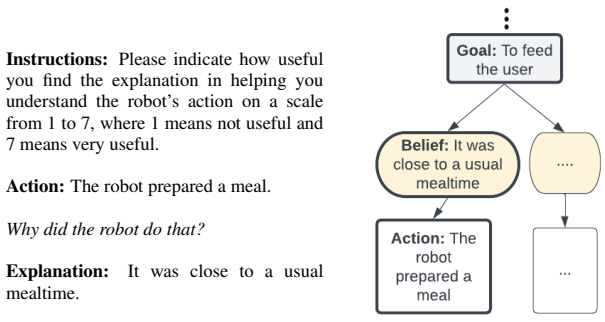
- H 3.2 Perception of causality (dispositionism vs. interactionism) is associated with explanation preference.
- H 3.3 Attention (field vs. focal objects) is associated with explanation preference.

4 Method

4.1 Instrument

The questionnaire administered for this study was divided into four parts: (1) Robot explanations, (2) Robot attitudes and perceptions, (3) Cognitive factors and (4) Socio-cultural information. The schematic structure of the questionnaire is represented in Figure 1.

Robot explanations. The first part of the questionnaire was designed to collect data for constructing the dependent variable, capturing participants' perceived usefulness for the robot's explanations. Participants were introduced to 28 robot actions along with one explanation per action, and they were asked to rate the explanation based on perceived usefulness.



(a) Generated Item

(b) An excerpt of the Goal Tree Hierarchy

Figure 2: Example of item (a) with a *belief explanation*, generated from the excerpt of the Goal Tree Hierarchy in (b).

The explanations were of four types: (1) *belief explanations*, (2) *goal explanations*, (3) *hybrid explanations*, which consisted of both the relevant goal and belief and (4) baseline explanations, which just restated the actions as the explanation.To generate the actions and explanations for types 1-3 (belief, goal, and hybrid explanations), we first created a goal tree hierarchy, that represents the robot's reasoning. A goal tree hierarchy is an established method for representing a robot's high-level reasoning comprising the possible robot's actions, goals and beliefs, and it can be extracted directly from architectures like the BDI agent architecture [Harbers *et al.*, 2010b]. According to this goal tree hierarchy, the robot adopts the goals necessary to achieve its ultimate goal ("To meet the user's need" for our robot). Which actions are taken to achieve the goal depends on the robot's beliefs — see further about tree hierarchies and associated explanation generation in [Harbers *et al.*, 2010b; Kaptein *et al.*, 2017]. From the goal tree hierarchy, we derived the actions with the corresponding beliefs and goals, which serve as the explanations in this study. Figure 2 shows an example of one questionnaire item with a belief explanation. All the questionnaire items for this part and a sample of the goal tree hierarchy used to generate the explanations are in the Supplementary Materials[2].

Robot acceptance and perceptions. In the second part, we elicit how people perceive the robot introduced to them in the first part and their acceptance and perception of the robot. In particular, perceived rationality, perceived autonomy, attitude toward robots and robot anxiety are examined. **Perceived autonomy** is adapted from [Harbers *et al.*, 2017] and assessed by using a single Likert-type item asking the respondents *How autonomous do you consider the robot?* on a 7-point scale ranging from *Not at all* to *Fully*. Similarly, we elicited the **perceived rationality** by asking respondents to indicate *How rational do you consider the robot?* on a 7-point scale with the end-points *Not at all* and *fully*. **Attitude towards the robot** and **robot anxiety** are two constructs used for estimating the Unified Theory of Acceptance and Use of Technology model (UTAUT) [Venkatesh *et al.*, 2003]. For our purposes, we used items from the relevant constructs from [Venkatesh *et al.*, 2003] (adapted items A1, AF2 and Affect1 for attitude) and [Heerink *et al.*, 2009] (all items for anxiety and ATT6 for attitude), who adapted the UTAUT model to the case of assistive robots.

Cognitive factors. The third part of the questionnaire was dedicated to cognitive factors, namely the need for cognition and two constructs from the Analysis-Holism Scale (AHS), which are the locus of attention and perception of causality [Choi *et al.*, 2007]. For the **Need for Cognition Scale**, we used the abbreviated 3-item version used by [Buttrick and et al, 2019]. For **Locus of attention** and **perception of causality**, which measure analytic and holistic cognitive styles, we use the scale proposed and validated by Martín-Fernández et al. [Martín-Fernández *et al.*, 2022] with 3 items per construct.

Socio-cultural information. In the final part, we asked participants about their gender, religious affiliation, level of education, subject of education and political orientation.

4.2 Procedure

This study was registered at the IRB of our Institution, and the instrument was administered via Prolific[3] in August 2023. First, participants were given an information sheet and were requested to indicate their consent to participate in the study. Consenting participants then proceeded to the first part of the study, the *robot explanation part*. In the *robot explanation part*, participants were randomly assigned to one of four blocks. Each of these blocks comprised all 28 robot actions, but each action was accompanied by a different explanation type, depending on the block. For example, Action *A* is explained by *goal explanation* in *Block 1*, *belief explanation* in *Block 2*, *hybrid explanation* in *Block 3*, and *baseline explanation* in *Block 4*. All blocks contain 7 instances of each explanation type. The action-explanation pairs were presented in random order to alleviate the order effect. We aimed to create a situation in which participants engage with each item in relative isolation, to mimic an authentic instance of human-robot interaction, which would consist of one action performed by the robot, followed by one explanation. To achieve this, each item was presented on a new page. Before administering the final questionnaire, we ran a pilot study in Prolific (N=55). After the feedback in the pilot, we clarified some parts of the questionnaire. For example, we added a clarification that the task is to evaluate the usefulness of the provided explanation, rather than the usefulness of the robot's action.

4.3 Data Quality & Participants

We employed three well-known data quality methods for online surveys. First, we recruited participants on Prolific with an approval rate over 95% achieved in previous studies, who completed at least 10 studies [Peer *et al.*, 2014]. Second, we included 3 attention checks [Hauser and Schwarz, 2016] in the survey, one in the *robot explanation* part, one in the *robot acceptance section* and one in the *socio-cultural factors* section, to ensure that attention checks are distributed evenly throughout the questionnaire. Third, to identify straightlining, which is the practice of giving the same answer in a

[2]https://osf.io/htxm7/?view_only=db740f78e3c54d0cbe185f460e5b1a78

[3]https://www.prolific.co/

battery of items [Kim *et al.*, 2019], we inspected three constructs containing a reversed item (robot attitude, need for cognition and locus of attention) for straight-lining using the simple nondifferential method [Kim *et al.*, 2019].

We recruited 512 participants using a non-proportional quota sampling method to ensure sufficient representation of all variables and socio-cultural factors, which are known to play a role in explanation preferences [Kopecka *et al.*, 2024]. Also, because some of the factors may be nationality-dependent, we focused on UK participants for this study (see associated limitations and future work in Section 6.2). The final sample comprised the data from 468 respondents due to the exclusion of 9 participants for failing one or more attention checks and 35 participants for straight-lining. The details of the demographic composition of the sample is in the supplementary materials.

5 Results

5.1 Raw Explanation Scores

First, we focus on the *raw explanation scores*, which are four Likert scores, one per explanation type, computed as a sum of all Likert items representing the given explanation type (belief, goal, hybrid and baseline).

RQ1: Overall Rating of Explanation Types

The first research question is concerned with the overall differences in how people rate different explanation types. We observed that hybrid explanations are preferred by most users (64%, including cases where the hybrid explanations score is equal to another explanation type), and the remaining 36% is distributed between the remaining three explanation types. To analyse the results, we compared the mean ranks between explanation type scores. Due to the scores being non-normally distributed[4], we used the non-parametric Friedman test. Results of the Friedman test showed that there is a statistically significant difference between the perceived usefulness of explanation and explanation type ($\chi^2(3) = 635.716, p < .001$). According to Dunn's pairwise post hoc tests with Bonferroni correction for multiple testing, all pairwise combinations of the four explanation type scores (belief, goal, hybrid and baseline scores) are significantly different from each other ($p < .001$), except for belief score ($Mdn = 38$) and goal score ($Mdn = 36$). The hybrid score was rated the highest ($Mdn = 42$), followed by belief and goal score, while baseline explanation ($Mdn = 20$) performed the worst, which means that our results confirm hypothesis H1.1. Figure 3 represents the distribution of explanation scores and medians.

RQ2: Robot Perception and Acceptance

To investigate the relationship between cognitive factors, acceptance and perception of the robot and explanation type rating, we fitted a multiple multivariate regression model (Model 1), where perception, acceptance of the robot and users' cognitive factors are predictors and the outcome variables are the four raw explanation scores. The results are in Table 1.

[4] $D_{belief}(468) = .083, p < .001$, $D_{goal}(468) = .069, p < .001$, $D_{hybrid}(468) = .083, p < .001$, $D_{base}(468) = .118, p < .001$

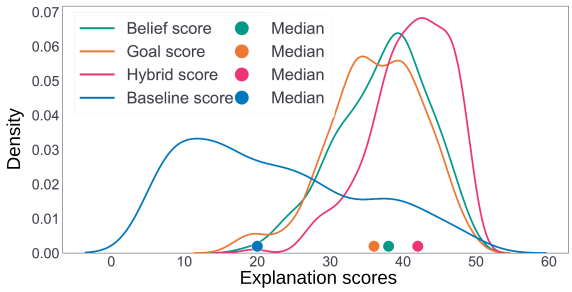


Figure 3: Density plots of the raw explanation scores.

The degree to which people find the domestic robot rational and autonomous is linked to how useful they find reason explanations in helping them to understand its actions. With increasing perception of the robot's rationality, people consider belief ($B = .837, p < 0.001$), goal ($B = .813, p = 0.001$) and hybrid explanations ($B = .601, p = 0.006$) more useful. As the perception of *autonomy* increases, belief ($B = .503, p = 0.041$) and hybrid explanations ($B = .459, p = 0.033$) are rated as more useful, but not goal explanations ($B = .221, p = 0.38$). While robot anxiety does not influence the perceived usefulness of explanations, attitude towards the robot does – the more positive attitude towards the robot one has, the higher are all explanations rated: belief ($B = .923, p = 0.002$), goal ($B = 1.427, p < 0.001$), hybrid ($B = .784, p = 0.003$) and baseline ($B = 1.559, p = 0.009$).

RQ3: Cognitive Factors

The only cognitive factor that shows a significant relationship with explanation types is the perception of causality ($B = .641, p = 0.011$). To remind the reader, perception of causality is a sub-construct of the Analysis-Holism scale, that measures one's tendency to attribute causality. Scoring high on the perception of causality measure indicates the tendency to consider complex interaction between the actor and the environment and when ascribing causality (interactionism), while low scores point to a tendency to ascribe causality to the actor (dispositionalism). Typically, interactionists consider a greater amount of information than dispositionists when attributing causes to behaviours [Choi *et al.*, 2007].

5.2 Adjusted Explanation Scores

The results of Model 1 reveal an interesting finding – some significant factors influencing perceived explanation usefulness have the same effect across explanation types (attitude towards the robot) and particularly across the reason explanation types (perception of rationality), rather than indicating a preference for a particular explanation type. To separate this tendency to rate all explanations higher, we calculated the *adjusted explanation scores (Δscores)*, which are derived from the raw explanation scores as the difference between the baseline explanation score and each reason explanation type (belief, goal and hybrid) per participant.

RQ1: Overall Rating of the Explanation Types

The distribution of the new scores Δ is depicted in Figure 4. Results of the Friedman test showed that there is a statistically

	Model 1 - *raw scores*				Model 2- *adjusted scores*		
	Belief	Goal	Hybrid	Base	ΔBelief	ΔGoal	ΔHybrid
Perception of rationality	0.837***	0.813**	0.601**	0.236	0.601	0.577	0.366
Perception of autonomy	0.503*	0.221	0.459*	-0.086	0.589	0.307	0.545
Robot anxiety	-0.343	-0.197	-0.413	1.021	-1.364*	-1.217*	-1.433*
Attitude towards robot	0.923**	1.427***	0.784**	1.559**	-0.636	-0.132	-0.775
Need for cognition	-0.535	0.092	0.093	-0.005	-0.53	0.096	0.097
Perception of causality	0.28	-0.046	0.641*	-0.841	1.122	0.796	1.482*
Locus of attention	-0.086	0.212	-0.119	1.085	-1.171*	-0.873	-1.204*
Intercept	30.127***	30.847***	35.346***	21.516***	8.611**	9.331**	13.83***

Table 1: We only show slope coefficients, full regression tables are available in supplementary materials.$p < .001 **$, $p < .01 **$, $p < .05*$

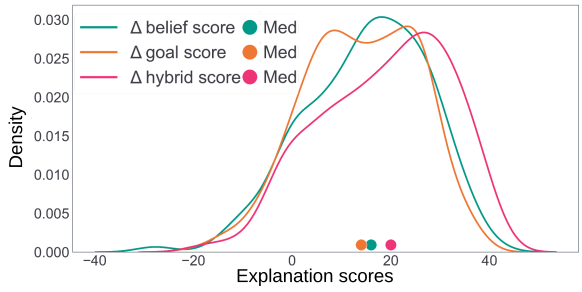


Figure 4: Density plots of the adjusted (Δ) explanation scores.

significant difference between the perceived usefulness of explanation and explanation type ($\chi^2(2) = 196.563, p < .001$). Similarly to the original scores, Δhybrid ($Mdn = 20$) is significantly better than Δbelief ($Mdn = 16$) and Δgoal ($Mdn = 14$) ($p < .001$), but in this case, Δbelief is rated higher than Δgoal, according to Dunn's pairwise post hoc tests with Bonferroni correction for multiple testing.

RQ2: Robot Acceptance and Perception

We fitted Model 2 (Table 1) to investigate the relationship between robot acceptance, perception and cognitive factors as predictors and Δbelief score, Δgoal score and Δhybrid score as outcome variables to observe the effect after accounting for the general tendency to like robot explanations observed earlier. A main difference from the results in Model 1 is the absence of a significant association between explanation scores and attitude towards the robot, and perceived rationality and autonomy of the robot. The interpretation of this could be that people with positive attitudes towards the robot (and to a lesser degree perceiving it as more rational and autonomous) indiscriminately liked all the explanations better and now that this aspect was subtracted from the scores, attitude towards the robot, perception of rationality and perception of autonomy no longer appear as significant predictors of the adjusted scores. However, we can observe robot anxiety, perception of causality and locus of attention to play a role in the perceived usefulness of explanations. Robot anxiety shows a significant negative association with all three adjusted scores: Δbelief ($B = -1.364, p = .018$), Δgoal ($B = -1.217, p = .023$) and Δhybrid ($B = -1.433, p = .018$)

RQ3: Cognitive Factors

Now, we proceed to examine the cognitive factors. According to our data, interactionism is associated with an increase in the perceived usefulness of hybrid explanations ($B = 1.482, p = .017$). The fact that interactionists consider both the actor and their environment in attributing causes to action seems consistent with our finding that they prefer to receive both goal and belief explanation (combined in hybrid explanation). Locus of attention is another cognitive factor associated with the rating of explanation types. This sub-construct of the Analysis-Holism scale describes the tendency to either focus on individual details (low LoA score) or the 'whole picture' (high LoA score). According to Model 2, those who are more oriented towards seeing the whole rather than the details are associated with a decrease in the Δbelief and Δhybrid.

6 Discussion

We found in our results support for several of the hypotheses tested — see the table in supplementary material for a complete summary. For instance, hybrid explanations are rated as the best (H1.1), perceived rationality and autonomy and high acceptance of the robot are associated with an increase in the usefulness of belief explanations (H2.1, H2.2) even though this effect is not unique to the belief explanation type, and we partially confirm that perception of causality and locus of attention is associated with a difference in explanation type (H3.2, H3.3).

6.1 Main Takeaways

1) Hybrid explanations are preferred by most users. We provide empirical evidence that explanations combining both goals and beliefs seemed to be preferred by most users to explanations consisting only of either of these elements and as such we validate and quantitatively confirm the recommendation suggested by [Harbers *et al.*, 2010a; Kopecka *et al.*, 2024]. Therefore, a recommendation for designing robot explanations is to use both goals and beliefs (which in some architectures like BDI agents may be readily available). Next, we offer several, more nuanced takeaways.

2) General trends are observed across explanation types. Some factors influence the perception of several or even all explanation types. The perceived rationality of the robot is connected with an increase in finding all the reason explanations more useful (not the baseline). Finding reasons

and not the baseline explanation more useful seems consistent with goals and beliefs (and their combination) being reasons and thus associated with rationality [Malle, 2011], and our results indicate that perceived rationality is indiscriminately linked with all the explanations consisting of reasons. Another factor that has a general effect across explanation types is an attitude towards the robot — it is linked to the increased usefulness of all explanation types (Model 1), which means that those with a positive attitude towards the robot consider all the explanation types more useful, perhaps indicating an overall tendency to be more positive about interacting with the robot and about all the explanations.

3) Some factors are associated with certain explanation types. For example, perceived robot autonomy is only associated with explanations containing beliefs, which are belief and hybrid explanations. The fact that the robot is capable of perceiving the environment (forming beliefs) and using this knowledge for reasoning and subsequently explaining its actions indicates the robot's ability to act autonomously in the environment. Users who perceive the robot as more autonomous might rank these types of explanations higher as they fit into the mental model they formed of the robot. Goal explanations, however, are not associated with the perception of autonomy. This could be because having a goal does not imply the capacity for autonomy – the robot could be simply *programmed* to pursue the given goal.

Regarding the perception of causality, interactionists have a higher preference for hybrid explanations. Interactionism is the tendency to consider the interaction of the person and their environment when ascribing causes for their action rather than focusing only on the person's dispositions [Choi *et al.*, 2007], which seems consistent with our results that interactionists prefer to receive both the goal reason, which *might* reveal something about the robot's dispositions as well as the belief reason, that often represents some contextual information of the robot.

4) Controlling for the overall tendency to rate explanations positively reveals additional insights. We attempted to discount the general tendency to like explanations by computing the adjusted scores by subtracting the baseline score. When considering the adjusted score, we can observe that robot anxiety contributes to lower ratings for all the reason explanations. A possible explanation of this phenomenon could be that people with higher robot anxiety did not want to engage with the robot and its reasoning and hence disliked explanations that provided them with additional information (the effect is the strongest with hybrid explanations). People with higher levels of robot anxiety are known to have a tendency to avoid talking to robots [Nomura *et al.*, 2008; De Graaf and Ben Allouch, 2013] and prefer robots acting autonomously to avoid interacting with them [Chanseau *et al.*, 2016]. This paper might also contribute to the understanding of how robot anxiety causes avoidant behaviour in terms of explanations, but further research is needed to evaluate this.

The adjusted scores also reveal distinct preferences according to the cognitive style in locus of attention. Regarding locus of attention, people with global attention seem not to find belief and hybrid explanations as useful as the goal ones. This might be because they have a tendency to attend to the entire field and they might more easily find deficiencies or missing information in explanations that contain beliefs, while goal explanations provide only what the robot is trying to achieve, which in some cases may feel more complete/correct. This hypothesis would need to be confirmed by future research.

6.2 Limitations and Future Research

Deploying an online survey allowed us to reach a high number of participants and enabled the use of inferential statistical methods with enough statistical power, but this meant participants did not have the opportunity to interact with the robots physically, so they could not observe and interpret the context of the action. This, however, had the advantage that we could isolate the effect of the user characteristics central to this study (cognitive tendencies, and acceptance and perception of the robot), as the context of the human-robot interaction and the interpretation of the action are known to affect explanation needs [Wachowiak *et al.*, 2023; Ferreira and Monteiro, 2020; Wachowiak *et al.*, 2024]. Further research should focus on investigating the possible interactions of the user characteristics studied in this paper and contextual factors. Another limitation of our study is that our participant sample was drawn only from the UK. This was done to avoid uncontrolled effects because national context is known to affect factors like perception of causality, locus of attention [Choi *et al.*, 2007], preferences and perceptions of robots [Lee and Sabanović, 2014; Lee *et al.*, 2012]. This means that the findings should not be extrapolated outside of the UK national context. Future work could replicate this study in different national contexts.

7 Conclusion

This paper examined the role of cognitive factors and the acceptance and perception of domestic service robots in the relationship with the perceived usefulness of four types of explanations that could be used by the robot to explain its action. The types of explanations examined were informed by folk-psychology and hence based on the same concepts people use to explain their behaviour. Despite identifying some differences based on the investigated human factors, a hybrid explanation consisting of the robot's goal and belief seems to be the most preferred explanation for explaining the robot's action. Additionally, this paper contributes to the understanding of more nuanced differences between users according to their cognitive factors and their robot acceptance and perception. We found that some user factors affect the perceived usefulness of all explanation types, while other factors have a discriminating effect between different kinds of explanations.

Ethical Statement

This study was registered at IRB of our institution as a Minimal Risk Study as the study posed no foreseeable risk to the participants and all collected data were anonymous. Hence, the authors do not foresee any ethical issues with this study.

Acknowledgements

We would like to thank anonymous reviewers, Munkhtulga Battogtokh, Sara Tandon and Lennart Wachowiak for their

feedback on the questionnaire and Zoe Evans on the robot explanations items. This work was supported by UK Research and Innovation [grant number EP/S023356/1], in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence (www.safeandtrustedai.org), and the INCIBE's strategic SPRINT (Seguridad y Privacidad en Sistemas con Inteligencia Artificial) C063/23 project with funds from the EU-NextGenerationEU through the Spanish government's Plan de Recuperación, Transformación y Resiliencia.

References

[Abdul *et al.*, 2018] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *CHI*, pages 1–18, 2018.

[Anderson *et al.*, 2021] Andrew Anderson, Tianyi Li, Mihaela Vorvoreanu, and Margaret Burnett. Human-ai interaction for diverse humans: What cognitive style disaggregation reveals. *arXiv preprint arXiv:2108.00588*, 2021.

[Buttrick and et al, 2019] Nicholas Buttrick and et al. Cross-Cultural Consistency and Relativity in the Enjoyment of Thinking Versus Doing. *Journal of Personality and Social Psychology*, 117(5):71–83, 2019.

[Cacioppo and Petty, 1982] John T. Cacioppo and Richard E. Petty. The need for cognition. *Journal of Personality and Social Psychology*, 42:116–131, 1982.

[Chanseau *et al.*, 2016] Adeline Chanseau, Kerstin Dautenhahn, Kheng Lee Koay, and Maha Salem. Who is in charge? sense of control and robot anxiety in human-robot interaction. In *2016 25th IEEE RO-MAN*, pages 743–748, 2016.

[Choi *et al.*, 2007] Incheol Choi, Minkyung Koo, and Jong An Choi. Individual differences in analytic versus holistic thinking. *Personality and Social Psychology Bulletin*, 33(5):691–705, 2007.

[Conati *et al.*, 2021] Cristina Conati, Oswald Barral, Vanessa Putnam, and Lea Rieger. Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial Intelligence*, 298:103503, 2021.

[De Graaf and Ben Allouch, 2013] Maartje M.A. De Graaf and Somaya Ben Allouch. The relation between people's attitude and anxiety towards robots in human-robot interaction. In *2013 IEEE RO-MAN*, pages 632–637, 2013.

[De Graaf and Malle, 2017] Maartje M.A. De Graaf and Bertram F. Malle. How people explain action (and autonomous intelligent systems should too). *AAAI Fall Symposium*, pages 19–26, 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. *ACM/IEEE HRI*, 2019-March:239–248, 2019.

[Ferreira and Monteiro, 2020] Juliana Ferreira and Mateus Monteiro. What Are People Doing About XAI User Experience? A Survey on AI Explainability Research and Practice. In *Design, User Experience, and Usability. Design for Contemporary Interactive Environments*, pages 56–73, Cham, 2020.

[Ferrer *et al.*, 2021] Xavier Ferrer, Tom van Nuenen, Jose Such, Mark Cote, and Natalia Criado. Bias and discrimination in ai: a cross-disciplinary perspective. *IEEE Technology and Society*, 20(2):72–80, 2021.

[Harbers *et al.*, 2010a] Maaike Harbers, Joost Broekens, Karel Van Den Bosch, and John Jules Meyer. Guidelines for developing explainable cognitive models. In *ICCM*, pages 85–90, 2010.

[Harbers *et al.*, 2010b] Maaike Harbers, Karel Van Den Bosch, and John Jules Meyer. Design and evaluation of explainable BDI agents. *IEEE/WIC/ACM IAT*, 2:125–132, 2010.

[Harbers *et al.*, 2017] Maaike Harbers, Marieke MM Peeters, and Mark A Neerincx. Perceived autonomy of robots: effects of appearance and context. In *A World with Robots: International Conference on Robot Ethics: ICRE 2015*, pages 19–33. Springer, 2017.

[Hauser and Schwarz, 2016] David J Hauser and Norbert Schwarz. Attentive turkers: Mturk participants perform better on online attention checks than do subject pool participants. *Behavior research methods*, 48:400–407, 2016.

[Heerink *et al.*, 2009] Marcel Heerink, Ben Kröse, Vanessa Evers, and Bob Wielinga. Measuring acceptance of an assistive social robot: A suggested toolkit. *IEEE International Workshop on Robot and Human Interactive Communication*, pages 528–533, 2009.

[Kaptein *et al.*, 2017] Frank Kaptein, Joost Broekens, Koen Hindriks, and Mark Neerincx. Personalised self-explanation by robots: The role of goals versus beliefs in robot-action explanation for children and adults. *RO-MAN*, pages 676–682, 2017.

[Kim *et al.*, 2019] Yujin Kim, Jennifer Dykema, John Stevenson, Penny Black, and D. Paul Moberg. Straightlining: Overview of measurement, comparison of indicators, and effects in mail–web mixed-mode surveys. *Social Science Computer Review*, 37(2):214–233, 2019.

[Kopecka *et al.*, 2024] Hana Kopecka, Jose Such, and Michael Luck. Preferences for ai explanations based on cognitive style and socio-cultural factors. *Procs. of the ACM on Human-Computer Interaction*, 8(CSCW1):1–32, 2024.

[Lee and Sabanović, 2014] Hee Rin Lee and Selma Sabanović. Culturally variable preferences for robot design and use in south korea, turkey, and the united states. In *HRI*, page 17–24. ACM, 2014.

[Lee *et al.*, 2012] Hee Rin Lee, JaYoung Sung, Selma Šabanović, and Joenghye Han. Cultural design of domestic robots: A study of user expectations in korea and the united states. In *IEEE RO-MAN*, pages 803–808, 2012.

[Malle and Knobe, 1997] Bertram F. Malle and Joshua Knobe. The folk concept of intentionality. *Journal of Experimental Social Psychology*, 33(2):101–121, 1997.

[Malle, 1999] Bertram F Malle. How People Explain Behavior: A New Theoretical Framework. *Personality and Social Psychology Review*, 3(1):23–48, 1999.

[Malle, 2011] Bertram F. Malle. Time to give up the dogmas of attribution. an alternative theory of behavior explanation. In *Advances in experimental social psychology*, volume 44, pages 297–352. Elsevier, 2011.

[Martín-Fernández *et al.*, 2022] Manuel Martín-Fernández, Blanca Requero, Xiaozhou Zhou, Dilney Gonçalves, and David Santos. Refinement of the Analysis-Holism Scale: A cross-cultural adaptation and validation of two shortened measures of analytic versus holistic thinking in Spain and the United States. *Personality and Individual Differences*, 186(February), 2022.

[Masuda and Nisbett, 2001] Takahiko Masuda and Richard E. Nisbett. Attending Holistically Versus Analytically: Comparing the Context Sensitivity of Japanese and Americans. *Journal of Personality and Social Psychology*, 81(5):922–9343, 2001.

[Millecamp *et al.*, 2019] Martijn Millecamp, Nyi Nyi Htun, Cristina Conati, and Katrien Verbert. To explain or not to explain: The effects of personal characteristics when explaining music recommendations. In *IUI*, page 397–407, 2019.

[Millecamp *et al.*, 2020] Martijn Millecamp, Robin Haveneers, and Katrien Verbert. Cogito ergo quid? the Effect of Cognitive Style in a Transparent Mobile Music Recommender System. In *UMAP*, pages 323–327, 2020.

[Miller, 2019] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[Mohseni *et al.*, 2018] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Trans. Interact. Intell. Syst.*, 11(3–4), aug 2018.

[Nomura *et al.*, 2008] Tatsuya Nomura, Takayuki Kanda, Tomohiro Suzuki, and Kensuke Kato. Prediction of human behavior in human–robot interaction using psychological scales for anxiety and negative attitudes toward robots. *IEEE Trans. on Robotics*, 24(2):442–451, 2008.

[Nosek *et al.*, 2018] Brian A Nosek, Charles R Ebersole, Alexander C DeHaven, and David T Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018.

[Nunes and Jannach, 2017] Ingrid Nunes and Dietmar Jannach. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction*, 27:393–444, 2017.

[Peer *et al.*, 2014] Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. Reputation as a sufficient condition for data quality on amazon mechanical turk. *Behavior research methods*, 46:1023–1031, 2014.

[Pugh, 2020] Jonathan Pugh. *Autonomy, rationality, and contemporary bioethics*. Oxford University Press, 2020.

[Rao and Georgeff, 1995] Anand Rao and Michael Georgeff. BDI Agents: From Theory to Practice. *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS-95)*, 95:312–319, 1995.

[Reich and Eyssel, 2013] Natalia Reich and Friederike Eyssel. Attitudes towards service robots in domestic environments: The role of personality characteristics, individual interests, and demographic variables. *Paladyn, Journal of Behavioral Robotics*, 4(2):123–130, 2013.

[Ribera and Lapedriza, 2019] Mireia Ribera and Àgata Lapedriza. Can we do better explanations? a proposal of user-centered explainable ai. In *IUI Workshops*, 2019.

[Robbemond *et al.*, 2022] Vincent Robbemond, Oana Inel, and Ujwal Gadiraju. Understanding the Role of Explanation Modality in AI-assisted Decision-making. *UMAP*, pages 223–233, 2022.

[Sanneman and Shah, 2022] Lindsay Sanneman and Julie A. Shah. The situation awareness framework for explainable ai (safe-ai) and human factors considerations for xai systems. *International Journal of Human–Computer Interaction*, 38(18-20):1772–1788, 2022.

[Song and Kim, 2022] Christina Soyoung Song and Youn-Kyung Kim. The role of the human-robot interaction in consumers' acceptance of humanoid retail service robots. *Journal of Business Research*, 146:489–503, 2022.

[Such, 2017] Jose Such. Privacy and autonomous systems. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4761–4767, 2017.

[van Nuenen *et al.*, 2020] Tom van Nuenen, Xavier Ferrer, Jose Such, and Mark Cote. Transparency for whom? assessing discriminatory artificial intelligence. *IEEE Computer*, 53:36–44, 2020.

[Varnum *et al.*, 2010] Michael E.W. Varnum, Igor Grossmann, Shinobu Kitayama, and Richard E. Nisbett. The origin of cultural differences in cognition: The social orientation hypothesis. *Current Directions in Psychological Science*, 19(1):9–13, 2010.

[Venkatesh *et al.*, 2003] Viswanath Venkatesh, Michael G. Morris, Gordon B. Davis, and Fred D. Davis. User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3):425–478, 2003.

[Wachowiak *et al.*, 2023] Lennart Wachowiak, Oya Celiktutan, Andrew Coles, and Gerard Canal. A survey of evaluation methods and metrics for explanations in human–robot interaction (hri). In *ICRA2023 Workshop on Explainable Robotics*, 2023.

[Wachowiak *et al.*, 2024] Lennart Wachowiak, Andrew Fenn, Haris Kamran, Andrew Coles, Oya Celiktutan, and Gerard Canal. When do people want an explanation from a robot? In *Proc. of the 2024 ACM/IEEE HRI*, 2024.

[Winikoff and Sidorenko, 2023] Michael Winikoff and Galina Sidorenko. Evaluating a mechanism for explaining bdi agent behaviour. In *EXTRAAMAS@AAMAS*, pages 18–37, 2023.