Contrastive Explanations of BDI Agents

Michael Winikoff
Victoria University of Wellington
Wellington, New Zealand
michael.winikoff@vuw.ac.nz

ABSTRACT

The ability of autonomous systems to provide explanations is important for supporting transparency and aiding the development of (appropriate) trust. Prior work has defined a mechanism for Belief-Desire-Intention (BDI) agents to be able to answer questions of the form "why did you do action X?". However, we know that we ask *contrastive* questions ("why did you do X *instead of* F?"). We therefore extend previous work to be able to answer such questions. A computational evaluation shows that using contrastive questions yields a significant reduction in explanation length. A human subject evaluation was conducted to assess whether such contrastive answers are preferred, and how well they support trust development and transparency. We found some evidence for contrastive answers being preferred, and some evidence that they led to higher trust, perceived understanding, and confidence in the system's correctness. We also evaluated the benefit of providing explanations at all. Surprisingly, there was not a clear benefit, and in some situations we found evidence that providing a (full) explanation was worse than not providing any explanation.

KEYWORDS

Explainable Agents; Contrastive explanations; Belief-Desire-Intention

ACM Reference Format:
Michael Winikoff. 2026. Contrastive Explanations of BDI Agents . In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 10 pages. https://doi.org/10.65109/VRTD1803

1 INTRODUCTION

Explainability of autonomous systems (e.g. [34, 35, 44]) is important for a number of reasons. These include supporting transparency [21, 23, 46], aiding the development of *appropriate* levels of trust [29, 39, 47], and a range of other reasons (e.g. acceptability [13], understandability [45], accountability [11], and traceability [45]). In particular, it is important to provide means of engineering autonomous systems that can explain their behaviour in human-meaningful terms [33, 40, 41, 50].

There is a whole body of work on explainable AI (XAI) [3]: techniques that allow explanations to be provided for the behaviour of AI modules. However, despite the importance of explainability of autonomous systems, most of the work on XAI has focused on explaining machine learning ("data-driven XAI" [3]), with a much smaller body of work focusing on explaining autonomous agents ("goal-driven XAI" [3], or "explainable agency" [29]).

Prior work exploring how humans explain themselves [31] has shown that in similar contexts (i.e. explaining the reasons for choosing a specific course of action) humans use the concepts of beliefs, desires, and *valuings*. Subsequently, Winikoff et al. [50] noted the natural correspondence between these concepts and the Belief-Desire-Intention (BDI) agent architecture [5, 6, 38], and defined a mechanism[1] that allows BDI agents to provide explanations of their actions in terms of these concepts, answering questions of the form "Why did you do action X?".

However, it is known that as humans we often ask *contrastive* questions of the form "Why did you do X (the *fact*) *instead of* F (the *foil*)?" (although sometimes the "instead of F" is implicit) [28, 33]. There is also empirical evidence that such explanations are most effective [37].

Although there has been some recent work on generating contrastive explanations[2] in different settings (e.g. planning [4, 9, 42, 43], rule-based systems [20], and Markov Decision Processes [2, 10]) and domains (e.g. driving [17, 37] and robotics [27, 36]), we are not aware of any work that addresses generating contrastive explanations of actions for BDI agents. The closest is Jasinski et al. [24] which focuses on contrastive explanations for the *selection* of goals, not the actions taken by the agent.

This paper makes three contributions: (1) extending Winikoff et al. [50] to generate contrastive explanations for BDI agents (§3); (2) conducting a *computational* evaluation (§4) to assess the extent of the reduction in explanation length; and (3) conducting a *human subject* evaluation (§5) to assess whether contrastive explanations are *preferred* and whether they are *effective*.

2 BACKGROUND

We firstly (§2.1) define goal-plan trees and then (§2.2) define (non-contrastive) explanation generation.

2.1 Goal-Plan Trees

A goal-plan tree is a standard abstraction of a wide range of BDI agent platforms, which specify agent behaviour using event-triggered plans with a context condition and plan body. A goal-plan tree is either an *action* node (which has no children), or a *goal* node (which has at least one child with each child having a condition). Where a node has multiple children we refer to them as "siblings", and to the

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/VRTD1803

[1] Earlier work by Harbers *et al.* [7, 18, 19] also considered BDI agents and defined a mechanism for providing explanations. However, they defined particular fixed patterns for extracting reasons from goal-plan trees, rather than a general algorithm, and did not make the link to the evidence on how humans explain themselves.

[2] These are related but distinct from counterfactuals: a contrastive question is asking "Why?" about something that was actually done (with reference to a possible alternative), whereas a counterfactual question is a *hypothetical* about something that was *not* done.

ones appearing earlier in the sequence as being "older". All nodes have a *name*. Action nodes also have a pre- and post-condition. Goal nodes also have a *type*. At a high level we have And (do all child nodes) and Or nodes (select one child node and do it). However, we also make a number of other distinctions, which lead us to use the following node types: *all* (all children are done in unconstrained order), *seq* (all children are done in sequential order), *one* (one child is selected and done), *sone* (one child is selected and done, but the selection process considers the children in a specified order, i.e. for a child to be selected its condition must be True and the conditions of all its "older" siblings must be False), and *xone* (one child is selected and done, but, for explanatory purposes, it is indicated that the children are mutually exclusive: there is no situation in which more than one child is available for selection). We use And to refer to *all* and *seq* and Or to refer to the *one*, *sone* and *xone* node types. Formally we define a goal tree GT as follows:

$$\begin{aligned} GT &::= (ActionName, Pre, Post) \mid (GoalName, Type, Child^{+}) \\ Child &::= (Cond, GT) \\ Type &::= all \mid seq \mid one \mid sone \mid xone \end{aligned}$$

We use as an example the coffee scenario from Winikoff et al. [50] (see [48, Figure 3]). The scenario is that the agent has the desire to getcoffee, and has three ways to do so: (i) get (low quality but free) coffee from a kitchen, which requires a staff card (either one's own, or a colleague's), and involves the sequence: getStaffCard (sub-goal), goto(kitchen), getCoffee(kitchen) (both actions); (ii) get (decent quality free) coffee from a colleague's office, which uses pods and requires the colleague, Ann, to be in their office, and involves the sequence of actions goto(office), getPod, getCoffee(office); or (iii) get (good but expensive) coffee from a shop, which requires money, and involves the sequence of actions goto(shop), pay(shop), getCoffee(shop).

2.2 Explanation Generation

We view an explanation as a set of time-tagged *explanatory factors* F, where each factor is either a *desire*, a *belief*, or a *valuing*. Formally: $F ::= \text{D:}_t N \mid \text{B:}_t C \mid \text{V:}_t N_1 < N_2$ where N is the name of a node, C is the condition of a node or a pre-condition of an action, and $N_1 < N_2$ indicates that N_2 is preferred over N_1, and t is a time tag which we omit in the remainder of the paper. We define $\mathcal{E}_X^T$ as being the set of explanatory factors that explain action X with regard to trace (i.e. sequence of actions) T and in the context of the given goal tree. We assume (not formalised below) that $X \in T$, otherwise the response to "why did you do X?" is "I didn't".

The basic idea is that to explain "why did you do X?" we consider the three types of explanatory factors. **(1) Desires:** What was the goal that was trying to be achieved? This can be found by considering the ancestors of X in the tree. However, we filter out Or nodes because they provide less information than their child, for example, "I want to get coffee" is less informative than "I want to get office coffee" (Figure 1, Line 1; ignore green shaded parts). **(2) Beliefs:** What conditions held (or failed to hold) that were relevant to choosing to do X? There are three cases here. The first two cases relate to why we were *able* to do X, whereas the third (and Valuings, below) relate to why we *chose* to do X. **Case 1:** Pre-conditions of X and of actions that were done before X (i.e. appear before it in the trace): these must hold, otherwise X could not be reached and done (Figure 1, Line 2). We filter these by removing pre-conditions of Y that were achieved by another action Z that necessarily occurs before Y in the trace (Figure 1, Line 9), e.g. getting shop coffee would not include the condition that we were at the shop (pre-condition of getCoffee(shop)) because it is a post-condition of goto(shop). **Case 2:** Conditions on nodes: let N be either X or an ancestor of X, then its condition had to hold for X to be reached, so is included in the explanation (Line 3). For example, I chose to go to the shop because I had money, so that option was available. **Case 3:** Excluded choices: when we choose a particular child of a *one* node, some of the other possible choices (siblings) are excluded because their conditions did not hold. This is part of the explanation for why we chose the child we did. For example, I chose to go to the shop because I could not get office coffee because Ann was not in her office. More precisely, let O be a node of type *one* that is an ancestor of X, and that has children $N_1, \ldots, N_x, \ldots N_n$ where X is a descendant of N_x or $X = N_x$. Then for those N_i ($i \neq x$) where their condition is known to not hold, this condition is also part of the explanation (Line 4). For *sone* this is modified (Line 6): we only include the conditions of siblings older than N_x (i.e. N_i where $i < x$). For *xone* we do not include the conditions of any siblings: the fact that the condition of N_x was True implies that the conditions of N_i ($i \neq x$) must be False. **(3) Valuings:** These also relate to choices. Where we chose a particular child of a *one* node, then the other options (siblings) that were available but weren't selected are less preferred. More precisely, let O be a node of type *one*[3] that is an ancestor of X with children $N_1, \ldots, N_x, \ldots N_n$ where X is a descendent of N_x or $X = N_x$. Then for nodes N_i ($i \neq x$) where the condition of N_i held, we chose to pursue N_x instead of N_i because N_x was more valued (Line 7)[4]

Figure 1 shows the formal definition of $\mathcal{E}_X^T$ (for now please ignore the parts that are boxed and shaded green). In addition to a number of straightforward auxiliary functions and predicates (noted in the figure's caption), we also define two predicates that are a little more complex. The first is $sib(N, X, N_x, N_i)$ (Line 10) which is true when N_i is a child of N that is a sibling of X or of N_x, where N_x is an ancestor of X. The second is $seqBef(N_1, N_2)$ (Line 11) which is true when N_1 necessarily occurs before N_2 in the trace (i.e. their common ancestor is a *seq* node and N_1 occurs before N_2 in the sequence).

Example: Returning to the coffee example, consider a situation where the agent performed the sequence of actions: getOwnCard, followed by goto(kitchen), and then getCoffee(kitchen). The answer to the (non-contrastive) question "why did you do getOwnCard?"—where we assume that a colleague's card is available and that one has money, but that Ann is not in her office—is the following set of six explanatory factors:

{ D:getKitchenCoffee, B:ownCard, B:staffCardAvailable,
B:¬AnnInOffice, V:getOthersCard<getOwnCard,
V:getShopCoffee<getKitchenCoffee }

[3] For sequential and exclusive Or we don't have more than one child as an option, so Valuings are not part of the explanation for those nodes.
[4] Adding value effects to the goal-plan tree can be done straightforwardly, following Winikoff et al. [50], and allows for more precise explanations for why N_x was valued more than N_i.

$$\mathcal{E}^T_{X\,/F} = \{\text{D:}N \mid ancest_F(N,X) \wedge \neg isOne(N) \wedge \neg isXOne(N) \wedge \neg isSOne(N)\} \quad (1)$$

$$\cup \{\text{B:}filter(A_i) \mid \lceil i \leq j \rceil \; i = j\} \quad (2)$$

$$\cup \{\text{B:}cond(N) \mid N = X \vee ancest_F(N,X)\} \quad (3)$$

$$\cup \{\text{B:}\neg cond(N_i) \mid ancest^{ca}_F(N,X) \wedge isOne(N) \wedge sib(N,X,\_,N_i) \wedge nheld(N_i)\} \quad (4)$$

$$\cup \{\text{B:}\neg cond(N_i) \mid ancest^{ca}_F(N,X) \wedge isSeqOne(N) \quad (5)$$

$$\wedge sib(N,X,N_x,N_i) \wedge sib(N,F,N_f,\_) \wedge seqn(N_f) \leq seqn(N_i) < seqn(N_x)\} \quad (6)$$

$$\cup \{\text{V:}N_i{<}N_x \mid ancest^{ca}_F(N,X) \wedge isOne(N) \wedge sib(N,X,N_x,N_i) \wedge held(N_i)\} \quad (7)$$

$$\text{Where } T = \langle A_1 \ldots A_j = X \ldots A_n \rangle \quad (8)$$

$$filter(A_i) = pre(A_i) \setminus \{c \mid seqBef(N,A_i) \wedge isAct(N) \wedge N \in T \wedge c \in post(N)\} \quad (9)$$

$$sib(N,X,N_x,N_i) = child(N_x,N) \wedge (ancest(N_x,X) \vee N_x = X) \wedge child(N_i,N) \wedge x \neq i \quad (10)$$

$$seqBef(N_1,N_2) = (ancest(N'_1,N_1) \vee N'_1 = N_1) \wedge (ancest(N'_2,N_2) \vee N'_2 = N_2) \quad (11)$$
$$\wedge parent(N'_1) = parent(N'_2) \wedge isSeq(parent(N'_1)) \wedge seqn(N'_1) < seqn(N'_2)$$

$$ancest_F(N,X) = ancest(N,X) \wedge \neg ancest(N,F) \quad (12)$$

$$ancest^{ca}_F(N,X) = ancest_F(N,X) \vee ca(N,X,F) \quad (13)$$

$$ca(C,A,B) = ancest(C,A) \wedge ancest(C,B) \wedge \neg\exists N.ancest(N,A) \wedge ancest(N,B) \wedge ancest(C,N) \quad (14)$$

$$\mathcal{E}^T_{X/?} = \bigcup_{f \in vf(X)} \mathcal{E}^T_{X/f} \quad (15)$$

$$first(N) = \begin{cases} \{N\}, & \text{if } N.type = Action \\ \bigcup_{N_i \in children(N)} first(N_i), & \text{if } N.type \neq seq \\ first(N_1), & \text{if } N.type = seq \wedge child(N_1,N) \wedge seqn(N_1) = 1 \end{cases} \quad (16)$$

$$vf(X) = \{F \mid ca(C,X,F) \wedge (isOne(C) \vee isXOne(C) \vee isSOne(C)) \wedge child(N_X,C) \wedge (ancest(N_X,X) \vee N_X = X) \wedge X \in first(N_X) \wedge child(N_F,C) \wedge (ancest(N_F,F) \vee N_F = F) \wedge F \in first(N_F)\} \quad (17)$$

Figure 1: Definition of $\mathcal{E}^T_X$ (ignore green shaded parts) and of $\mathcal{E}^T_{X/F}$ (adding green shading). Auxiliary functions and predicates used: $pre(N)$ and $post(N)$ (return pre-/post-condition of an action as a conjunction represented as a set of propositions), $cond(N)$ (returns the condition of node N), $parent(N)$ (return the parent node of N in the given goal tree), $children(N)$ (return the children of N), $seqn(N)$ (returns the sequence number of N), $ancest(A,B)$ (true when A is an ancestor of B), $child(N_i,N)$ (true when N_i is a child of N), $isOne(N)$ (resp. $isXOne$, $isSOne$, and $isSeq$) which is true when the type of node N is one (resp. $sone$, $xone$, seq), and $isAct(N)$ (true when N is an action node). We also use the predicate $held(N)$ which is true when the condition of node N held when the node was reached (or if there is no condition), and $nheld(N)$ which is true when the condition of node N did not hold.

3 CONTRASTIVE EXPLANATION GENERATION

In defining contrastive explanations we begin with the observation that a contrastive query is in effect a *filter*. When asked "why did you do X and not F?" we are still explaining why we did X, but the explanation is focussed on those things that explain specifically why not F, in other words we filter out things that don't specifically relate to explaining why not F. We next review each of the three explanatory factors to consider how the explanation generation needs to be modified to do this filtering. We then consider how to handle implicit contrastive questions, which requires us to first define what is a valid foil.

Desires: recall that these are non-Or ancestor nodes of the query X. However, any node that is an ancestor of X and also an ancestor of F can be filtered out, because it does not relate to explaining why not F. We define this using a modified *ancest* predicate (denoted $ancest_F$) that is only true for ancestors of X that are not also ancestors of F (Figure 1, Line 12). Line 1 of Figure 1 is then modified to

use $ancest_F$ instead of *ancest*. **Beliefs:** we have three cases. **Case 1:** Pre-conditions of X and of actions that were done before X: the pre-conditions of X remain relevant. However, the pre-conditions of earlier actions are not relevant. Specifically, following Grice's maxims [15], we argue that by asking the question "why X and not F?" the asker is indicating that X is the earliest action which was unexpected. Hence, we assume that actions preceding X are correct and not relevant. We formalise this by replacing $i \leq j$ (indicated with a dashed box in Line 2) with $i = j$. **Case 2:** Conditions of ancestor nodes: Similarly to the case for Desires, we omit ancestor nodes of X that are also ancestor nodes of F (Line 3 uses $ancest_F$ instead of *ancest*). **Case 3:** Choice-related non-holding conditions: Similarly to the case for desires, we focus on ancestors that are not also ancestors of F, except that we do want to also consider the closest common ancestor, because this is the point where a decision was made that related to the choice between X and F. We therefore (Lines 4 & 6) use $ancest^{ca}_F$ instead of *ancest*, where $ancest^{ca}_F(N, X)$ is defined (Line 13) as ancestors of X that are not also ancestors of F, but including their closest common ancestor (identified using $ca(C, A, B)$, Line 14). Additionally, for *sone* we also filter out explanatory factors associated with N_i that appear earlier than N_f: since they appear before N_f, they cannot relate to the difference between doing X and doing F (Line 6, $seqn(N_f) \leq seqn(N_i)$). **Valuings:** Similarly to the earlier cases, we exclude nodes that are also ancestors of F. This is done (Line 7) by using $ancest^{ca}_F$ instead of *ancest*. Figure 1 shows the formal definition of the explanatory factors for "why did you do X instead of F?" (given tree G and trace T) (denoted $\mathcal{E}^T_{X/F}$).

Example: Returning to the coffee example with the same sequence of actions as in §2.2, the answer to the contrastive question "why did you do getOwnCard instead of getOthersCard?" ($\mathcal{E}^T_{\text{getOwnCard/getOthersCard}}$) is the following set of two explanatory factors: {B:ownCard, V:getOthersCard<getOwnCard}.

Implicit foils: As noted earlier, sometimes a contrastive question has an *implicit foil*, i.e. the foil is implied rather than explicitly stated. We handle this by identifying all possible foils, and then taking the weakest (least restrictive) filter: we only filter out an explanatory factor if it is filtered out for all possible foils. This corresponds to taking the *set union* of the explanations for each of the foils: if a factor is in an explanation for any foil, it is included. We therefore define the answer to an implicit contrastive question, denoted $\mathcal{E}^T_{X/?}$, as the union of $\mathcal{E}^T_{X/f}$ over the set of possible valid foils $vf(X)$ (defined below), yielding Line 15 in Figure 1.

Valid foils: We now need to consider the question of what is a valid foil and define $vf(X)$. Intuitively, a valid foil F is one that was not done (i.e. $F \notin T$), but could have been done, replacing the fact X. Unfortunately implementing this intuition would require exploring possible traces to check that there exists one where F is done and replaces X. We therefore instead use the condition that the fact X and prospective foil F have a common ancestor node C that is an Or node, and that X and F are both possible first actions (Figure 1 Line 16) of the relevant child of C. This is sufficient to ensure the desired intuitive conditions are met (see Line 17).

4 COMPUTATIONAL EVALUATION

It is clear that contrastive explanations are shorter, but not by how much. We assess the length reduction by generating 1000 random trees of sufficient size ($\geq \theta$) and comparing the size of non-contrastive and contrastive explanations for each action node N in the tree and for each of its valid foils. Since the explanation generation function $\mathcal{E}$ requires a trace T, we also generate a trace for each N (see §4.2). Pseudo-code summarising the process can be found in [48, §A.1].

4.1 Tree Generation

We follow the "traditional" BDI goal-plan tree structure which alternates And and Or nodes, and only permits children of Or nodes to have conditions. This means that an Or node has non-Or nodes as children, so if the children of an Or (i.e. *one*, *sone* or *xone*) are not action nodes, then they must be either *all* or *seq* (selected randomly). Similarly, the children of And nodes (i.e. *seq* or *all*) are either action nodes or Or nodes (with type selected randomly from *one*, *sone*, *xone*). A generated tree is either an action node (with probability α, as long as it is at depth $d > 1$), or a goal node with n children ($2 \leq n \leq \epsilon$, where ϵ is a parameter of the tree generation) and a type (one of *one*, *sone*, *xone*, *seq*, *all*). Children of a *sone* or *seq* node have a sequence number. All action nodes have a pre-condition, and each node that is the child of an Or node has a condition. We limit the depth of the tree to δ: if the depth is reached during the tree generation process then we only generate children that are action nodes. Python code implementing tree generation is in [48, §A.2] along with brief discussion of how our tree generation differs from Yao and Wu [51].

4.2 Trace Generation

We need to be able to generate a trace from a given tree $\mathcal{R}$ and a selected node $N \in \mathcal{R}$. In the process of generating the trace, we assign truth values to conditions in a way that is consistent with N being selected. This assignment is needed because the explanation generation uses these (*held*, *nheld*). We therefore define a procedure generate_trace(R, N), where R is a node (initially the root of tree $\mathcal{R}$), that returns a trace and annotates nodes in $\mathcal{R}$ to indicate whether their conditions hold.

The process of generating a trace is recursive, starting with the root of the tree. If we reach an action node A then we mark the node's condition as being True and return the trace $\langle A \rangle$. Otherwise, we proceed as follows, depending on the type of R, which we assume has children $N_1 \ldots N_n$.

If R is an Or node then we first select the node N_x ($1 \leq x \leq n$) such that N_x is either equal to the node N for which an explanation will be generated, or is an ancestor of N. We then mark the condition of N_x as True. For R of type *one* we mark the conditions of sibling nodes N_i ($i \neq x$) randomly (nodes without conditions can only be marked True). For R of type *xone* we mark the conditions of N_i as False (following the exclusive-or semantics of *xor*), and for R of type *sone* we mark nodes N_i ($i < x$) as False (following the sequential-or semantics of *sor*). We then return the trace generated from N_x (generate_trace(N_x, N)).

If R is an And node then we generate the traces T_i recursively for each N_i, and then concatenate these traces sequentially (for R

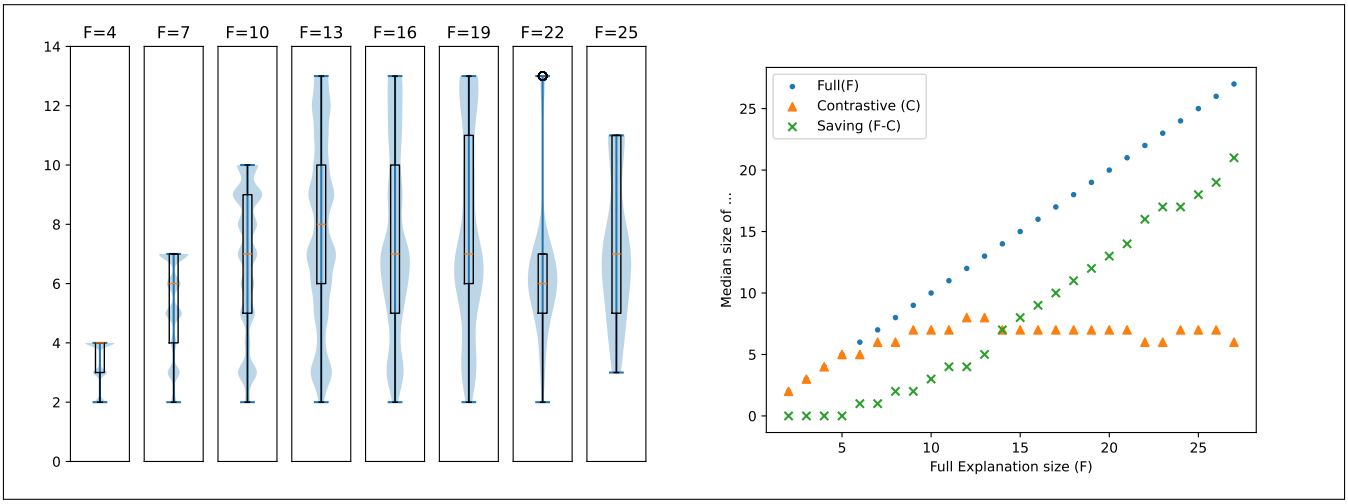

Figure 2: Computational Evaluation Results: plot of distribution of contrastive size against full explanation size for selected F (left) and of median full size / contrastive size / saving against full explanation size (right).

of type *seq*) or interleave them randomly (for R of type *all*). We use $\fatsemi$ to denote the function that joins two traces in sequence and $|||$ to denote the function that randomly interleaves two traces.

```
def generate_trace(R, N):
  if isAct(N) then
    mark(N,True)
    return ⟨N⟩
  if isOr(R) then // find N_x
    for N_i ∈ children(R):
      if ancest(N_i, N) ∨ N_i = N then N_x = N_i
    mark(N_x, True)
    for N_i ∈ (children(R) \ {N_x}):
      if isOne(R) then markRandom(N_i)
      if isXOne(R) then mark(N_i,False)
      if isSOne(R) ∧ seqn(N_i)<seqn(N_x) then
        mark(N_i, False)
    return generate_trace(N_x, N)
  if isAnd(R) then
    for i ∈ {1...n}: T_i ← generate_trace(N_i, N)
    if isAll(N) then return T_1 ||| ... ||| T_n
    if isSeq(N) then return T_1 ⨟ ... ⨟ T_n
```

4.3 Results

We generated 1000 trees using the following generation parameter values: $\alpha = 0.5$ (probability of a node being an action node), $\delta = 5$ (maximum tree depth), $\epsilon = 5$ (maximum number of children), and $\theta = 20$ (minimum tree size).

These parameter values are based on the goal-plan tree of Broekens et al. [7]. This tree has 22 nodes (consistent with $\theta = 20$), and each non-leaf node has 2-5 children (hence $\epsilon = 5$). The tree has 14 actions, of which 9 are at the bottom of the tree (depth of 4), so considering nodes that are *not* at the maximum depth, for which α applies, we have 5 action nodes and 7 non-action nodes (4 at depth 3 and 3 at depth 2), which is consistent with $\alpha = 0.5$. Finally, this tree has a maximum depth of 4. However, it does not follow the traditional BDI structure of alternating Or and And, and we therefore increase δ to 5, since having an Or node with an Or node as a child, say, would require an additional intervening And node in the traditional BDI structure.

Figure 2 shows the evaluation results. We use F to refer to the size of the Full (non-contrastive) explanation and C to refer to the size of the Contrastive explanation. Since the potential saving in size ($F - C$) is limited by the size of F, we plot (left side of Figure) the distribution of $F - C$ for different selected values of F. This shows that, especially for larger F, the saving by using contrastive explanations can be significant. We also plot the median values of F, C and $F - C$ for each value of F (right side of Figure). This shows that as the original explanation size (F, x-axis) increases beyond 10, the median size of the contrastive explanation (C, "▲") remains roughly fixed, and hence the saving ($F - C$, "x") grows.

Additional results can be found in [48, §A.3], including results for two example trees from the Intention Progression Competition[5], which also show substantial length reductions (C vs. F).

5 HUMAN SUBJECT EVALUATION

We have shown that contrastive explanations are significantly shorter. We would therefore expect that contrastive explanations would have lower cognitive load, since they are shorter, and hence be both *preferred* and more *effective*. To assess this we conduct a human subject evaluation, specifically answering the two questions: (1) are contrastive explanations *preferred*? (2) are they *effective* (at supporting appropriate trust and transparency)? These are distinct questions: people may prefer explanations that are less effective (e.g. [2]).

There has been a range of work on evaluating explanations [1, 2, 7, 16, 19, 25, 26, 37, 49], but none of it addresses our research

[5] https://gitlab.com/intention-progression-competition/example-gpts/

questions, mostly because contrastive explanations were not considered. Omeiza et al. [37] found that in the domain of autonomous vehicles participants given contrastive explanations did not have a significant difference in understanding compared to those receiving "Why" explanations, but that they were significantly better at assigning accountability for a collision or near miss.

5.1 Research Hypotheses

We have 9 hypotheses. H1-H3 relate to the *preference* between contrastive and full explanations, H4-H6 relate to the *effectiveness* of explanations, H7-H8 concern the difference between providing and not providing explanations at all, and H9 looks at the relationship between trust in the specific system at hand and general trust in technology. We hypothesise that, since people tend to ask contrastive questions, and these give shorter more focused answers, that these would be preferred to full explanations (**H1**), have higher *perceived quality* (**H2**), be more likely to be seen as having the right level of detail (**H3**), yield higher trust (**H4**), higher belief in understanding of the system (**H5**), and higher confidence in the system's correct behaviour (**H6**). We also hypothesise that providing an explanation (either contrastive or full) yields higher trust (**H7**) and more confidence in the system's correct behaviour (**H8**) than not providing an explanation. Finally, we hypothesise (**H9**) that there is a correlation between trust in technology in general and trust in each of the two systems, but that the strength of the correlation is not high.

5.2 Methodology

We recruited participants[6] using the Prolific platform, requesting a gender-balanced sample of adults (18 years or older) who are fluent in English. Each participant was randomly allocated in a balanced way to one of three groups (**X**): FULL explanation, CONTrastive explanation or no explanation ("NONE"), and asked to complete a survey.

The survey (see [48, §B]) had the following sections: **(1) Consent. (2) Technology Trust (TT):** we measured the participant's general trust in technology, adopting the questions used by Winikoff and Sidorenko [49], which are based on Mcknight et al. [32], responses were on a 7 point Likert scale. **(3) System presentation:** Participants were presented with a high-level description of a system (see [48, §A.4]): the first system was a robot making pancakes (based on Broekens et al. [7]), the second was a search-and-rescue unmanned aerial vehicle (UAV). **(4) Scenario presentation:** For each system, three scenarios for that system were presented ([48, §A.4]), each giving the current situation, indicating what the system did, and then giving no explanation or a full explanation or a contrastive explanation[7]. **(5) Explanation Quality (Q):** For each scenario, participants who received an explanation were asked questions about the quality of the explanation, adopting the instrument from Hoffman et al. [22, Table 3], as well as a question about understanding (**U**) following Buçinca et al. [8], and about the level of detail (**LD**) of the explanation being too little, about right, or too much. Responses (Q & U) were on a 5 point Likert scale. **(6) Trust (Short):** For each scenario, all participants were asked about their trust in the system ("short trust"). Because this question is repeated for each scenario, we only asked one question here: "*I trust the autonomous system*" (response on a 7 point Likert scale). Participants were also asked about their belief in the correctness (**COR**) of the system: "*Do you believe that the system did what it should in this case?*" (possible responses: Yes, No, Unsure). **(7) Trust (Long):** After the third scenario for the system was presented, and associated questions answered (i.e. (5) & (6)), trust in the system ($\mathbf{T}_{Pan}$ and $\mathbf{T}_{SAR}$) was assessed, following Hoffman et al. [22, Table 8]. Answers were on a 5 point Likert scale. **(8) Preferred Explanation (PRE):** For each of the 6 scenarios, the participant was given a description of the situation and the system's selected course of action, and then asked whether they preferred the full explanation, the contrastive explanation, or whether they preferred the explanations equally. All participants were asked the same questions in this section. Responses were on a 5 point Likert scale (1 = "I strongly prefer explanation 1", 2 = "I somewhat prefer explanation 1", 3 = "I prefer these explanations equally", 4 = "I somewhat prefer explanation 2", 5 = "I strongly prefer explanation 2"; explanation 1 was the full explanation, and explanation 2 was the contrastive one). **(9) Demographic information:** We used the same questions as Sidorenko & Winikoff [49], asking about age, highest level of education, gender, ethnicity, and programming experience. With the two systems, and three scenarios for each system, the sequence of survey sections that participants saw was: 12 34564564567 34564564567 89 (participants in the NONE group were not shown section 5 so instead saw: 12 34646467 34646467 89).

5.3 Results

We now describe our results. We begin (§5.3.1) with a summary of the responses received, and the checking and filtering that was done, as well as demographics. We then proceed to present the results, beginning with an analysis of participants' *preferences* between contrastive and full explanations (§5.3.2, H1-H3), and then the *effectiveness* of explanations (e.g. resulting trust) comparing contrastive and full explanations (§5.3.3, H4-H6). We then consider the difference between having and not having explanations (§5.3.4, H7-H8), and finally we consider the relationship between trust in each of the two autonomous systems and general trust in technology (§5.3.5, H9), before finishing with discussion of the results (§5.3.6). Table 1 has a summary of the hypotheses, variables, tests used, and the results.

5.3.1 Responses and Filtering. We received 161 responses. These were filtered by removing low-quality responses using two mechanisms. Firstly, the survey included two attention check questions, which asked for agreement with a statement that was clearly false. Any participant who failed both attention checks was blocked from completing the survey. Participants who failed exactly one of the questions were able to complete (as per Prolific's policy on attention check questions), but their responses were not used in the analysis. Out of the 161 responses, 12 failed both attention check questions, and 18 failed exactly one of the questions. This left 131 responses. Secondly, the survey included for each scenario a short trust question, and a longer set of questions measuring trust for each system. A number of participants had inconsistent responses

[6]The experiment had ethics approval from the human ethics committee of Victoria University of Wellington (approval number 2024/HE000068).
[7]The explanations were generated by software from a goal-plan tree and manually rendered in English.

Hyp	Variables	Test	Results	Section
H1	PRE	1SW	✔ (partly, see text)	§5.3.2
H2	X_2-Q	M-W	✘	
H3	X_2-LD	M-W	✘	
H4	X_2-T_X	M-W	✔	§5.3.3
H5	X_2-U	M-W	✔ (scenario 2 only)	
H6	X_2-COR	M-W	✔ (scenarios 1&2 only)	
H7	X-T_X	K-W	✘	§5.3.4
H8	X-COR	K-W	✘ (see text)	
H9	TT-T_X	SRC	✔	§5.3.5

Table 1: Summary of Hypotheses, Variables, Tests, and Results. Key for variables: X = explanation (full, contrastive, or none), X_2 (only full and contrastive), T_x = trust in system $x \in \{Pan, SAR\}$, PRE = preference for explanation type, Q = quality of explanation, U = belief in understanding of system, COR = belief the system did the right thing, LD = level of detail of explanation (too little, about right, too much), TT = technology trust. Key for tests: M-W = Mann-Whitney, K-W = Kruskal-Wallis, 1SW = one-sample Wilcoxon, SRC = Spearman's rank correlation.

for these questions. We calculated a difference score by normalising each trust measurement to be out of 10 and then considered the difference to be too big if it was 2 or more out of 10. This check was done separately for each of the systems (pancake and search-and-rescue) and resulted in the exclusion of a further 27 participants, leaving 104 responses for analysis. We also manually checked the responses of participants who had completed the survey unusually quickly, but these all appeared reasonable.

For the three variables that were measured by multiple questions we calculated Cronbach's α to check that they were internally consistent. All were high enough ($\alpha \geq 0.8$).

The demographic profile of the 104 responses is as follows. Gender: 50 Male, 54 Female. Age: 23 participants were aged 18-24, 49 were aged 25-34, 15 (35-44), 10 (45-54), 5 (55-64), 2 (65-74), 0 (75+). Education: 22 (completed high school), 56 (completed undergrad degree), 23 (Masters), 2 (PhD), 1 (declined to answer). Ethnicity: 40 (African), 36 (European), 8 (North American), 7 (South American), 5 (Asian), 3 (Other), 2 (Australian), 2 (declined), 1 (New Zealander). Programming experience: 38 (None), 28 (hobby), 12 (studied at high school), 12 (currently studying degree), 10 (completed degree), 4 (other). An analysis of the variables vs. demographic factors found a few differences[8], but no age-related differences were found.

5.3.2 Explanation Type Preferences (H1-H3). Hypothesis 1 (contrastive explanations are preferred to full) is **partially confirmed**: for some scenarios there was a statistically significant preference for contrastive explanations, but for others a preference for full explanations. Specifically, a one sample Wilcoxon signed rank test of the preference variable (PREF) for each scenario (looking for a difference to not having a preference, i.e. value of 3) found statistically significant differences for scenarios 1-4 (respectively with $p = 0.001596$, $p = 0.0001423$, $p = 0.003559$, $p = 0.000008994$) and no difference for scenarios 5 and 6 ($p = 0.2914$ and $p = 0.8142$). For scenarios 1 and 2 there was a preference for contrastive explanations (median=4). However, for scenarios 3 and 4 there was a preference for full explanations (median=2). Scenario 5 had a (non-statistically significant) preference for contrastive explanations, and scenario 6 had a bimodal distribution (See [48, Figure 5] for the distribution of responses for scenarios 3, 5 and 6).

Hypothesis 2 (contrastive explanations have higher perceived quality) is **not confirmed**. A Mann-Whitney test comparing the quality of explanation scores (dependent variable Q) against the explanation type (X_2) did not find a statistically significant difference for any of the scenarios (p values for the six scenarios were respectively: 0.1525, 0.2246, 1.0, 0.7868, 0.805 and 0.6825).

Hypothesis 3 (contrastive explanations are more likely to be considered to have a level of detail (LD) that is "about right") is **not confirmed**. A Mann-Whitney test comparing the assessment of the level of detail (dependent variable LD) against the explanation type (X_2) did not find any statistically significant differences (p for the scenarios respectively 0.83, 0.45, 0.10, 0.95, 0.53, 0.47).

5.3.3 Effects of Explanation Type (H4-H6). Hypothesis 4 (H4) is that contrastive explanations yield a higher level of trust than full explanations. This hypothesis is **confirmed**: a Mann-Whitney test comparing the dependent variable of trust in the given system (T_{Pan} and T_{SAR}) against the independent variable of explanation type (X_2) shows a statistically significant difference for both systems with $p = 0.008575$ for the Pancake system and $p = 0.04205$ for the search-and-rescue system. On the five-point Likert scale where higher means more trusted, the median score for contrastive explanations was 3.5 for the Pancake system and 3.67 for search-and-rescue, and for full explanations was 3.0 for Pancake and 3.17 for search-and-rescue.

Hypothesis 5 (H5) is similar to H4 but with respect to the (single) question about the participants' perceived understanding of the system (U). This hypothesis was **confirmed**, but only for Scenario 2. A Mann-Whitney test comparing the understanding scores (dependent variable U) against the explanation type (X_2) found a statistically significant difference for scenario 2 only (p values for the six scenarios were respectively: 0.1193, 0.04079, 0.2715, 0.3987, 0.3663, and 0.7789).

Hypothesis 6 (H6) is that contrastive explanations yield more confidence in the system's correct behaviour (COR) than full explanations. This hypothesis is **confirmed**, but only for two of the scenarios. A Mann-Whitney test comparing confidence in the system's correct behaviour (dependent variable COR) against the explanation type (X_2) found $p = 0.005351$ for scenario 1 and $p = 0.04028$ for scenario 2 (p values for the other scenarios: 0.06866, 0.07875, 0.07907, and 0.9511). For scenario 1 the mean responses for the contrastive and full explanation groups were respectively 2.852941 and 2.342857. For scenario 2 these were 2.852941 and 2.514286.

5.3.4 Effects of not having Explanations (H7-H8). Hypothesis 7 (H7) is that both types of explanation yield higher trust than no explanation. This hypothesis is **not confirmed**: a Kruskal-Wallis test with pairwise Dunn test using the Holm correction method for multiple tests found no significant difference in trust (dependent variable T_x) between either contrastive or full compared with not

[8] Males had a higher T_{SAR} (p=0.01057, median 3.5 vs. 3.0 for females), a lower U for scenario 2 (p=0.00455, mean 3.545 vs. 4.306), and a lower Q for scenario 2 (p=0.04198, median 3.43 vs. 3.79). African participants had a higher TT than Europeans (p=0.01559249, median 5.86 vs. 4.91).

providing any explanation (X). Interestingly, the median trust when no explanation was provided was actually *higher* than the trust when a full explanation was provided (3.33 compared with 3.17 for the Pancake system, and 3.5 compared with 3 for the search-and-rescue system), but this difference was not statistically significant.

Hypothesis 8 (H8) is that having an explanation yields more confidence in the system's correct behaviour than not having an explanation. This hypothesis is **not confirmed**. In fact, there was a statistically significant difference in the *opposite* direction: a Kruskal-Wallis test with pairwise Dunn test using the Holm correction method for multiple tests found a significant difference between the confidence in the system's correctness (dependent variable COR) and the explanation type (X) for two of the scenarios ($p = 0.01046$ and 0.0124 respectively for scenarios 3 and 4) comparing full and no explanation, with the group that had been given full explanations having *lower* responses for the system's correctness. In other words, participants who were not given an explanation at all were more positive about the system's correctness than participants who had been given full explanations (no statistically significant difference was found for contrastive vs. no explanation).

5.3.5 Relationship between Trust in Technology and Trust in each system (H9). Hypothesis 9 is that there is a significant but medium-strength correlation between trust in a specific system (pancake robot or search-and-rescue, T_x) and trust in technology more generally (TT). This hypothesis is **confirmed**. For both specific systems, the Spearman's rank coefficient showed a statistically significant correlation ($p = 0.0000000153$ for the pancake robot and $p = 0.000065$ for search-and-rescue). For the pancake robot $\rho = 0.52$ is interpreted as a moderate strength relationship. For search-and-rescue $\rho = 0.38$ would be interpreted as either a moderate or weak strength relationship, depending on the boundaries used[9]. Graphs showing trust in each system against trust in technology in general can be found in [48, Figure 6]. Our result is consistent with the finding of Winikoff and Sidorenko [49] that trust in technology in general influences, but does not determine, trust in a particular system.

5.3.6 Discussion. There are a number of interesting points to draw out of these results. Firstly, we saw a difference between preference and effectiveness: contrastive explanations were not consistently preferred to full explanations, but they did give higher trust in the (trustworthy) system. The evaluation by Amitai et al. [2] also found this same outcome (albeit in a different setting). Secondly, we found that providing full explanations *reduced* trust in the system compared to not having any explanations. Kaptein et al. [26] similarly found that providing explanations resulted in participants being *less likely* to follow their system's recommendation. They speculated that providing information that participants already knew might reduce their adoption of the system's recommendations. We speculate that providing an explanation that is either too long or overly complex may result in a decrease in trust and confidence.

We also, similarly to prior evaluations [7, 19, 25, 37], found that results were sometimes scenario-dependent. One particular issue that may explain the scenario-dependent results for preferences is that for contrastive explanations the explanation is given with respect to a foil F. However, if the participant did not consider F to be a likely course of action, then the contrastive explanation might not match their expectations, and therefore not be preferred. This appears to be a plausible explanation for scenario 3 where the contrastive explanation used flipping the pancake by throwing it as foil, rather than using the spatula, and therefore the contrastive explanation filtered out the explanatory factor that the pancake was ready to be flipped.

[9]Sources vary in their definition of "moderate strength", e.g. 0.4 – 0.7 (https://tinyurl.com/3vct2nuj), or 0.3 – 0.7 (https://tinyurl.com/mr2jshkx) or 0.38 – 0.68 (https://tinyurl.com/2uc7wxcz).

6 CONCLUSION

We extended the prior work of Winikoff et al. [50] with contrastive explanations, which are well motivated in the literature [28, 33]. Our computational evaluation showed that contrastive explanations were significantly shorter. In particular, as the size of the (full) explanation grows the median size of the contrastive explanation does not grow, making it scalable. Our human subject evaluation found some (scenario-dependent) evidence of preference for contrastive explanations, and stronger evidence for the *effectiveness* of contrastive explanations (higher trust, better (self-assessed) understanding, and confidence in the system's correctness). We also found that participants who had not been given explanations had a *higher* level of confidence in the correct behaviour of the system for some scenarios than participants given full explanations.

Practical Implications: Firstly, when using contrastive explanations it is important to ensure that the foil matches the user's expectation. This can be done by having the user specify the foil explicitly (e.g. "why did you do X instead of F?"). Secondly, providing explanations is not risk-free. Poor quality or too-long explanations may actually reduce trust in the system. Therefore caution needs to be taken in deploying explanation facilities. Finally, human behaviour is complex. To avoid counter-intuitive results it is important to guide development and deployment with (carefully designed) user evaluations, and to involve representative participants in an iterative development process.

Future work includes further evaluation, with different scenarios and different systems, since our evaluation only used two simple scenarios. Another direction for future work is making explanations interactive, which can be done following a dialogue model [12], or as a graphical user interface. Another direction is to extend with additional *hypothetical* question types [30] such as: "what-if?" (what would happen if the situation was changed in a certain way), "how to be?" (what would need to change to obtain a certain behaviour), and "how to still be?" (what changes would leave the behaviour unchanged). Finally, this work could perhaps be made applicable to non-BDI agents by using policy graphs to model the observed behaviour of agents, and ascribing beliefs and intentions to these agents. Gimenez-Abalos et al. [14] have done this for "why?" questions.

ACKNOWLEDGMENTS

Michael Winikoff would like to thank the University of Ljubljana for hosting him on sabbatical while this paper was written.

REFERENCES

[1] Amal Abdulrahman, Deborah Richards, and Ayse Aysin Bilgin. 2022. Exploring the influence of a user-specific explainable virtual advisor on health behaviour change intentions. *Auton. Agents Multi Agent Syst.* 36, 1 (2022), 25. https://doi.org/10.1007/s10458-022-09553-x

[2] Yotam Amitai, Yael Septon, and Ofra Amir. 2024. Explaining Reinforcement Learning Agents through Counterfactual Action Outcomes. In *AAAI*. AAAI Press, 10003–10011. https://doi.org/10.1609/AAAI.V38I9.28863

[3] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable Agents and Robots: Results from a Systematic Literature Review. In *AAMAS*. 1078–1088. http://dl.acm.org/citation.cfm?id=3331806

[4] Rita Borgo, Michael Cashmore, and Daniele Magazzeni. 2018. Towards Providing Explanations for AI Planner Decisions. arXiv:1810.06338 [cs.AI] https://arxiv.org/abs/1810.06338

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

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

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

[8] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *International Conference on Intelligent User Interfaces (IUI)*. ACM, 454–464. https://doi.org/10.1145/3377325.3377498

[9] Michael Cashmore, Anna Collins, Benjamin Krarup, Senka Krivic, Daniele Magazzeni, and David Smith. 2019. Towards Explainable AI Planning as a Service. arXiv:1908.05059 [cs.AI] https://arxiv.org/abs/1908.05059

[10] Shenghui Chen, Kayla Boggess, and Lu Feng. 2020. Towards Transparent Robotic Planning via Contrastive Explanations. In *IROS* (Las Vegas, NV, USA). IEEE Press, 6593–6598. https://doi.org/10.1109/IROS45743.2020.9341773

[11] Stephen Cranefield, Nir Oren, and Wamberto Weber Vasconcelos. 2018. Accountability for Practical Reasoning Agents. In *Agreement Technologies (AT) (LNCS, Vol. 11327)*, Marin Lujak (Ed.). Springer, 33–48. https://doi.org/10.1007/978-3-030-17294-7_3

[12] Louise A. Dennis and Nir Oren. 2022. Explaining BDI agent behaviour through dialogue. *Auton. Agents Multi Agent Syst.* 36, 1 (2022), 29. https://doi.org/10.1007/S10458-022-09556-8

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

[14] Victor Gimenez-Abalos, Sergio Álvarez-Napagao, Adrián Tormos, Ulises Cortés, and Javier Vázquez-Salceda. 2024. Intention-aware policy graphs: answering what, how, and why in opaque agents. *CoRR* abs/2409.19038 (2024). https://doi.org/10.48550/ARXIV.2409.19038 arXiv:2409.19038

[15] H Paul Grice. 1975. Logic and conversation. In *Syntax and Semantics volume 3: Speech Acts*, P. Cole and J. Morgan (Eds.). Academic Press, New York.

[16] Balint Gyevnar, Stephanie Droop, Tadeg Quillien, Shay B. Cohen, Neil R. Bramley, Christopher G. Lucas, and Stefano V. Albrecht. 2025. People Attribute Purpose to Autonomous Vehicles When Explaining Their Behavior: Insights from Cognitive Science for Explainable AI. In *Conference on Human Factors in Computing (CHI)*. ACM, 86:1–86:18. https://doi.org/10.1145/3706598.3713509

[17] Balint Gyevnar, Massimiliano Tamborski, Cheng Wang, Christopher G. Lucas, Shay B. Cohen, and Stefano V. Albrecht. 2022. A Human-Centric Method for Generating Causal Explanations in Natural Language for Autonomous Vehicle Motion Planning. *arXiv* (2022). https://doi.org/10.48550/ARXIV.2206.08783

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

[19] Maaike Harbers, Karel van den Bosch, and John-Jules Ch. Meyer. 2010. Design and Evaluation of Explainable BDI Agents. In *Intelligent Agent Technology (IAT)*. IEEE, 125–132. https://doi.org/10.1109/WI-IAT.2010.115

[20] Lars Herbold, Mersedeh Sadeghi, and Andreas Vogelsang. 2024. Generating Context-Aware Contrastive Explanations in Rule-based Systems. In *Proceedings of the 2024 Workshop on Explainability Engineering (ExEn)*. 8–14. https://doi.org/10.1145/3648505.3648507

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

[22] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2023. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers Comput. Sci.* 5 (2023). https://doi.org/10.3389/FCOMP.2023.1096257

[23] IEEE. 2022. IEEE Standard for Transparency of Autonomous Systems. IEEE Std 7001-2021. , 54 pages. https://doi.org/10.1109/IEEESTD.2022.9726144

[24] Henrique Jasinski, Mariela Morveli-Espinoza, and Cesar Augusto Tacla. 2023. Towards Generating P-Contrastive Explanations for Goal Selection in Extended-BDI Agents. In *Intelligent Systems: 12th Brazilian Conference (BRACIS)*. 351–366. https://doi.org/10.1007/978-3-031-45368-7_23

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

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

[27] Raj Korpan, Susan L. Epstein, Anoop Aroor, and Gil Dekel. 2017. WHY: Natural Explanations from a Robot Navigator. arXiv:1709.09741 [cs.AI] https://arxiv.org/abs/1709.09741

[28] Benjamin Krarup, Senka Krivic, Daniele Magazzeni, Derek Long, Michael Cashmore, and David E. Smith. 2021. Contrastive Explanations of Plans through Model Restrictions. *J. Artif. Intell. Res.* 72 (2021), 533–612. https://doi.org/10.1613/JAIR.1.12813

[29] Pat Langley, Ben Meadows, Mohan Sridharan, and Dongkyu Choi. 2017. Explainable Agency for Intelligent Autonomous Systems. In *AAAI Conference on Artificial Intelligence*. AAAI Press, 4762–4764. http://aaai.org/ocs/index.php/IAAI/IAAI17/paper/view/15046

[30] Q. Vera Liao, Daniel M. Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Conference on Human Factors in Computing Systems (CHI)*. ACM, 1–15. https://doi.org/10.1145/3313831.3376590

[31] Bertram F. Malle. 2004. *How the Mind Explains Behavior*. MIT Press. ISBN: 9780262134453.

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

[33] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* 267 (2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[34] J. Müller and K. Fischer. 2014. Application Impact of Multi-agent Systems and Technologies: A Survey. In *Agent-Oriented Software Engineering*. Springer Berlin Heidelberg, 27–53. http://dx.doi.org/10.1007/978-3-642-54432-3_3

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

[36] Alberto Olivares-Alarcos, Gerard Canal, Sergi Foix, and Guillem Alenyà. 2024. Ontological modeling and reasoning for comparison and contrastive explanation of robot plans (extended abstract). In *Autonomous Agents and MultiAgent Systems (AAMAS)*. 2405–2407. https://www.ifaamas.org/Proceedings/aamas2024/pdfs/p2405.pdf

[37] Daniel Omeiza, Helena Web, Marina Jirotka, and Lars Kunze. 2021. Towards Accountability: Providing Intelligible Explanations in Autonomous Driving. In *2021 IEEE Intelligent Vehicles Symposium (IV)*. 231–237. https://doi.org/10.1109/IV48863.2021.9575917

[38] Anand S. Rao and Michael P. Georgeff. 1992. An Abstract Architecture for Rational Agents. In *Knowledge Representation and Reasoning (KR)*. Morgan Kaufmann Publishers, San Mateo, CA, 439–449.

[39] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M. Howard, and Alan R. Wagner. 2016. Overtrust of Robots in Emergency Evacuation Scenarios. In *Human Robot Interaction (HRI)*. IEEE/ACM, 101–108. https://doi.org/10.1109/HRI.2016.7451740

[40] Sebastian Rodriguez and John Thangarajah. 2024. Explainable Agents (XAg) by Design (Blue Sky Ideas Track). In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, N. Alechina, V. Dignum, M. Dastani, and J.S. Sichman (Eds.). ACM.

[41] Sebastian Rodriguez, John Thangarajah, and Andrew Davey. 2024. Design Patterns for Explainable Agents (XAg). In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, N. Alechina, V. Dignum, M. Dastani, and J.S. Sichman (Eds.). ACM.

[42] Mir Sarwar, Rajarshi Ray, and Ansuman Banerjee. 2023. A Contrastive Plan Explanation Framework for Hybrid System Models. *ACM Trans. Embed. Comput. Syst.* 22, 2, Article 22 (Jan. 2023), 51 pages. https://doi.org/10.1145/3561532

[43] Sarath Sreedharan, Siddharth Srivastava, and Subbarao Kambhampati. 2021. Using state abstractions to compute personalized contrastive explanations for AI agent behavior. *Artif. Intell.* 301 (2021), 103570. https://doi.org/10.1016/j.artint.2021.103570

[44] M. Birna van Riemsdijk, Catholijn M. Jonker, and Victor R. Lesser. 2015. Creating Socially Adaptive Electronic Partners: Interaction, Reasoning and Ethical Challenges. In *Autonomous Agents and Multiagent Systems (AAMAS)*. ACM, 1201–1206. http://dl.acm.org/citation.cfm?id=2773303

[45] Ruben S. Verhagen, Mark A. Neerincx, and Myrthe L. Tielman. 2021. A Two-Dimensional Explanation Framework to Classify AI as Incomprehensible, Interpretable, or Understandable. In *Explainable and Transparent AI and Multi-Agent Systems (EXTRAAMAS) (LNCS, Vol. 12688)*. Springer, 119–138. https://doi.org/10.1007/978-3-030-82017-6_8

[46] Alan F. T. Winfield, Serena Booth, Louise A. Dennis, Takashi Egawa, Helen F. Hastie, Naomi Jacobs, Roderick I. Muttram, Joanna I. Olszewska, Fahimeh Rajabiyazdi, Andreas Theodorou, Mark A. Underwood, Robert H. Wortham, and Eleanor Nell Watson. 2021. IEEE P7001: A Proposed Standard on Transparency. *Frontiers Robotics AI* 8 (2021), 665729. https://doi.org/10.3389/frobt.2021.665729

[47] Michael Winikoff. 2017. Towards Trusting Autonomous Systems. In *Engineering Multi-Agent Systems (EMAS) (LNCS, Vol. 10738)*. Springer, 3–20. https://doi.org/10.1007/978-3-319-91899-0_1

[48] Michael Winikoff. 2026. Contrastive explanations of BDI agents. https://doi.org/10.5281/zenodo.18603362

[49] Michael Winikoff and Galina Sidorenko. 2023. Evaluating a Mechanism for Explaining BDI Agent Behaviour. In *Explainable and Transparent AI and Multi-Agent Systems (EXTRAAMAS), Revised Selected Papers (LNCS, Vol. 14127)*. Springer, 18–37. https://doi.org/10.1007/978-3-031-40878-6_2

[50] Michael Winikoff, Galina Sidorenko, Virginia Dignum, and Frank Dignum. 2021. Why bad coffee? Explaining BDI agent behaviour with valuings. *Artif. Intell.* 300 (2021), 103554. https://doi.org/10.1016/J.ARTINT.2021.103554

[51] Yuan Yao and Di Wu. 2021. GenGPT: A Systematic Way to Generate Synthetic Goal-Plan Trees. In *Engineering Multi-Agent Systems (EMAS) (LNCS, Vol. 13190)*. Springer, 373–380. https://doi.org/10.1007/978-3-030-97457-2_21