Kaspar Causally Explains

H. Araujo[1], P. Holthaus[2], M, Sarda Gou[2], G, Lakatos[2], G. Galizia[2], L. Wood[2], B. Robins[2], M.R. Mousavi[1], and F. Amirabdollahian[2]

[1] King's College London, UK
[2] University of Hertfordshire, UK

Abstract. The Kaspar robot has been used with great success to work as an education and social mediator with children with autism spectrum disorder. Enabling the robot to automatically generate causal explanations is key to enrich the interaction scenarios for children and promote trust in the robot. We present a theory of causal explanation to be embedded in Kaspar. Based on this theory, we build a causal model and an analysis method to calculate causal explanations. We implement our method in Java with inputs provided by a human operator. This model automatically generates the causal explanation that are then spoken by Kaspar. We validate our explanations for user satisfaction in an empirical evaluation.

1 Introduction

Causality has intrigued philosophers since ancient times. Modern theories of causality were put forward by philosophers such as Hume and Lewis, and they found applications in engineering, particularly in explaining the results of testing and verification [1].

A formal logical theory that has been proved useful in engineering practice is the theory of actual causality by Halpern and Pearl [7]. This theory specifies the environment as a set of variables and a set of structural equations that describe the relation between them. Then, given a logical effect (represented as a Boolean predicate on the variables), a potential cause (also a Boolean predicate) through analysing counter-factuals in the causal model, i.e., parallel worlds in which the cause and effect may or may not have occurred.

This theory of causality has been proven useful both in analysing the results of testing and verification [2], as well as in providing explanations for complex systems such as neural networks [5]. In this paper, we apply Halpern and Pearl's theory of actual causality to provide explanations in educational scenarios for children in the autistic spectrum disorder. In particular, we equip Kaspar, which is a state-of-the-art humanoid robot primed for interaction with children with autism spectrum disorder (ASD) [15], with causal explanations. Enabling Kaspar to generate causal explanations is considered a key to enrich the interaction scenarios for children and thereby could promote additional trust in the robot.

As the main contribution of this work, we implement a tool that automatically builds a causal model and conducts a causal analysis to provide explanations behind certain events during interaction between children and Kaspar.

We implement our method in a Java implementation that, given inputs from a human operator, automatically generates the causal explanation that are spoken by Kaspar. Finally, we validate our explanations for user satisfaction in an empirical evaluation with healthy adults.

The rest of this paper is structured as follows. In Section 2, we discuss related work. In Section 3, we present an overview of Kaspar and the approach to its interaction with children. In Section 4, we present the discrete theory of actual causality by Halpern and Pearl. In Sections 5 and 6, we apply the theory to our context and present the mechanisation of our strategy, respectively. In Section 7, we present the explanations provided by Kaspar and our validation experiment. Lastly, in Section 8, we provide our conclusions and discuss future work.

2 Related Work

Our context is the theory of actual causality [7], where, in a given scenario leading to an outcome, the events are analysed in order to find causes. This is in contrast with type-level causality where general causal rules governing a system are sought. To our knowledge, no theory of causality has been applied to interactions with robots. Baier et al. [1] have conducted a survey on published approaches that utilise Halpern's notion of causality. We summarise some of the most prominent related work below.

Leitner-Fischer and Leue [11] define a theory of causality that considers the temporal order as well as the non-occurrence of events. They also provide a search-based on-the-fly causality assessment that does not require the counterexamples to be generated in advance. In our work however, the order of events does not play a role: whether a particular event occurred before or after other events does not impact on what we consider cause.

Beer et al. [2] use causal analysis to explain counterexamples in hardware verification. The proposed algorithm is implemented in the IBM RuleBase PE tool. Also, Chockler, Grumberg, and Yadgar [3] employ a notion of responsibility (degree of causality) [4] to improve the quality of abstraction refinement by producing mode efficient counterexamples. Besides the continuous aspects, our approach incorporates the modelling of platform (hardware), controllers (software) and environment into a single model that considers a high-level abstraction of the system. Considering a notion of responsibility is one of the directions for our future work to rank the explanations provided.

3 Application domain: Kaspar

The Kaspar robot [13] has been used to work with children with autism to help break their social isolation by acting as a social mediator with great success. A skill that children with ASD often struggle with is visual perspective taking (VPT), which is the ability to see the world from another person's viewpoint, making use of both spatial and social information [8]. Robot-mediated

intervention has already been shown to be efficient in teaching autistic children perspective-taking skills [10]. Enabling the robot to automatically generate causal explanations in such scenarios could be key to enrich the interaction scenarios for autistic children [14] and thereby promote additional trust in the education medium. This may in turn make the robot-mediation more successful.

With this purpose in mind, we designed 4 VPT interactive games for the children to play with Kaspar (discussed in Section 5.1). In these games, Kaspar asks the child to show a particular interactive object. In each of the scenarios, when Kaspar cannot see the animal that he has requested, he explains the reason why he cannot see it; for example: "I cannot see it because you are holding it too high" or "this is not the animal I have asked to see." By doing that, we expect the children to understand Kaspar's point of view and show the animals to Kaspar in the correct way.

4 Theory of Causal Explanation

In this section, we present Halpern and Pearl's theory of actual causality [7] and demonstrate its application through the following running example; the example is to illustrate the theory and the actual VPT games will follow.

Example 1. Consider a simple scenario where the Kaspar robot and a picture of a lion are sitting on a table. The picture is in Kaspar's line of sight. However, two things prevent Kaspar from actually seeing the lion: (i) Kaspar's eyes are shut (due to the press of a button by the teacher) and (ii) the picture falls off the table (due to the wind blowing from an open window). In this simple scenario, if Kaspar's eyes are closed then this is a cause for Kaspar not seeing the lion. Analogously, if the picture has fallen off the table, that is also a cause.

Mathematical assessments of causality require formal modelling. As a precondition to a model, a signature provides the set of variables and their admissible valuations. The formal definitions in this section are taken from those by Chockler and Halpern [4].

Definition 1 (Signature). *A signature is a tuple*

$$\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R}),$$

where $\mathcal{U}$ is a finite set of exogenous variables, $\mathcal{V}$ is a finite set of endogenous variables, and $\mathcal{R}$ associates with every variable $Y \in \mathcal{U} \cup \mathcal{V}$ a finite and nonempty set $\mathcal{R}(Y)$ of possible values for Y.

Exogenous variables are determined by factors outside of the model while endogenous variables are affected by exogenous ones and also by other endogenous variables. For instance, going back to the example, the state of Kaspar's eyes and the state of the picture can be seen as *endogenous* variables, but the presence of lighting that allows for Kaspar to see at all, whether the button was pressed, and whether the window is open are *exogenous* variable.

Example 2. A signature of our running example has the following variables. Below, we describe the variables for each set ($\mathcal{U}$ and $\mathcal{V}$) and their possible values that form $\mathcal{R}$.

In the exogenous set $\mathcal{U}$, we have that:

- u_b represents the button: 0 if it is not pressed and 1 if it is.
- u_w represents the window: 0 it is closed and 1 if it is open.

As for the endogenous set $\mathcal{V}$, we have:

- KE for Kaspar's eyes: 0 if they are open, and 1 if they are shut.
- PS for the picture: 0 if it is on the table and 1 if it has fallen off.
- KS for Kaspar's sight: 1 if Kaspar can see the lion and 0 if it cannot.

Definition 2 (Causal Model). *A causal model over a signature $\mathcal{S}$ is a tuple*

$$M = (\mathcal{S}, \mathcal{F}),$$

where $\mathcal{F}$ associates with each variable $X \in \mathcal{V}$ a function denoted by F_X, such that:

$$F_X : (\times_{U \in \mathcal{U}} \mathcal{R}(U)) \times (\times_{Y \in \mathcal{V} \setminus \{X\}} \mathcal{R}(Y)) \rightarrow \mathcal{R}(X)$$

F_X describes how the value of the endogenous variable X is determined by the values of all other variables in $\mathcal{U} \cup \mathcal{V}$. The indexed Cartesian products $\times_{U \in \mathcal{U}} \mathcal{R}(U)$ and $\times_{Y \in \mathcal{V} \setminus \{X\}} \mathcal{R}(Y)$ consider each possible values of the variables in $\mathcal{U}$ and $\mathcal{V} \setminus \{X\}$, respectively.

As mentioned, the set $\mathcal{U}$ of exogenous variables includes things we need to assume so as to render all relationships deterministic (such as the presence of light, wind conditions and whether someone has pressed the button). We denote $\vec{u}$ (i.e., a set of valuations in $\mathcal{R}(\mathcal{U})$) as the context behind a cause. That is, the context is a mapping of exogenous values to their variables that induce the value of the endogenous variables. Typically, the context can define the value of certain endogenous variables, which, in conjunction with the functions in $\mathcal{F}$, are used to determine the value of the remaining endogenous variables.

Consider that, in our example, the context comprises the unmodelled variables $u_w \in \mathcal{U}$ and $u_b \in \mathcal{U}$. They represent whether the window is open and whether someone has pressed the button that control Kaspar's eyes, respectively. These variables can be seen as inputs that are not controlled by the system.

Example 3. Given the signature of our running example (see Example 2), a causal model for this system can be defined with the following structural equations $\mathcal{F}$.

- $F_{KE}(\vec{u}, PS, KS) = u_b$
- $F_{PS}(\vec{u}, KE, PS) = u_w$
- $F_{KS}(\vec{u}, KE, PS) = 1 - max(KE, PS)$

In summary, the context affects whether Kaspar's eyes are shut or whether the picture is still on the table (i.e., KE and PS, respectively). Then, these variables affect whether Kaspar can see the lion (KS). As defined in the functions ($\mathcal{F}$) above, Kaspar can only see the lion if both KE and PS are 0.

Finally, to make the definition of cause precise, we first need a syntax for causal events. Given a signature $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$, a formula of the form $X = x$, for $X \in \mathcal{V}$ and $x \in \mathcal{R}(X)$, is called a primitive event.

Definition 3 (Causal Formula). *A causal formula is of the form*

$$[X_1 \leftarrow x_1, ..., X_k \leftarrow x_k]\Phi, \textit{where}$$

- $X_1, ..., X_k$ *are distinct variables in* $\mathcal{V}$.
- $x_i \in \mathcal{R}(X_i)$. *And,*
- Φ *is a Boolean combination of primitive events.*

The formula $[X_1 \leftarrow x_1, ..., X_k \leftarrow x_k]\Phi$ states that Φ holds in a system where X_i is set to x_i for $i = 1, ..., k$. Such a formula can be abbreviated as $[\vec{X} \leftarrow \vec{x}]\Phi$. An assignment of the type $X \leftarrow x$ (called an *intervention* by Halpern) can be interpreted as an update in $\mathcal{F}$ where the function for X is set just to x. In our Kaspar example, a valid causal formula is $[KE \leftarrow 1](KS = 0)$. This says that if Kaspar's eyes have been shut, then Kaspar cannot see the picture.

Thus, given a context $\vec{u} \in \mathcal{R}(\mathcal{U})$, we write $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}](Y = y)$ if the variable $Y \in \mathcal{V}$ has the value y in a causal model M where X_i is set to x_i for $i = 1, ..., k$. The notation can also be used in the presence of a Boolean combination of primitive events: $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}]\Phi$. Furthermore, in the special case where $k = 0$, we write $(M, \vec{u}) \models (Y = y)$ if the variable $Y \in \mathcal{V}$ has the value y given the context $\vec{u} \in \mathcal{R}(\mathcal{U})$ and the causal model M. This notation can also be used in the presence of a Boolean combination of primitive events: $(M, \vec{u}) \models \Phi$.

The types of events that are allowed as causes are of the form $(X_1 = x_1 \wedge ... \wedge X_k = x_k)$, that is, a conjunction of primitive events that can be abbreviated as $\vec{X} = \vec{x}$. Then, cause is formally defined as follows.

Definition 4 (Cause). *We say that* $\vec{X} = \vec{x}$ *is a cause of* Φ *in* $(M, \vec{u})$ *if the following three conditions hold:*

- *AC1.* $(M, \vec{u}) \models (\vec{X} = \vec{x}) \wedge \Phi$
- *AC2. There exists a partition* $(\vec{Z}, \vec{W})$ *of* $\mathcal{V}$ *with* $\vec{X} \subseteq \vec{Z}$ *and some setting* $(\vec{x'}, \vec{w'})$ *of the variables in* $(\vec{X}, \vec{W})$*, such that*
  - *a)* $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x'}, \vec{W} \leftarrow \vec{w'}]\neg\Phi$ *and,*
  - *b)* $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W'} \leftarrow \vec{w'}]\Phi$ *for all subsets* $\vec{W'}$ *of* $\vec{W}$.
- *AC3.* $(\vec{X} = \vec{x})$ *is minimal, that is, no subset of* $\vec{X}$ *satisfies AC2.*

Statement AC1 checks whether X and Φ are true at the same time, i.e., the cause has actually led to the effect; AC2 examines counterfactual dependence, i.e., given contingencies, changing other variables while keeping X intact brings about Φ and vice versa, changing X removes the effect, and AC3 checks whether

everything in X is actually necessary for Φ to be true (that is, whether X is minimal). An important aspect of this definition, which is also relevant to our application, is that the set of endogenous variables is split in two disjoint sets $\vec{W}$ and $\vec{Z}$, where $\vec{X} \subseteq \vec{Z}$. The variables in the set $\vec{W}$ allow for the cause to be tested under certain circumstances (called structural contingencies [4]) where the variables in $\vec{W}$ are set to $\vec{w}'$. The set $\vec{Z}$, however, comprises the variables that mediate the situation that makes Φ hold, when $\vec{X} \leftarrow \vec{x}$. That is, changing the values of variables in $\vec{X}$ might result in changing the values of other variables (i.e., $\vec{Z} - \vec{X}$), which then leads to Φ.

Example 4. Consider the scenario in our Kaspar example where $KE = 1, PS = 1, KS = 0$. We would like to assess whether the fact that the picture has fallen off ($PS = 1$) is the cause of Kaspar not seeing the lion ($KS = 0$).

AC1 states that $(\vec{X} = \vec{x})$ cannot be a cause of Φ, unless both the primitive causal events $(\vec{X} = \vec{x})$ and the effect Φ are true in the causal model M, given the context $\vec{u}$. That is, it states that for $PS = 1$ to be the cause of $KS = 0$, then both need to be true in $(M, \vec{u})$; thus, in this scenario, AC1 holds. Conversely, if we were trying to assess whether $KE = 0$ is the cause of $KS = 0$ then, AC1 could not be satisfied, as $KE = 0$ is not true in $(M, \vec{u})$ and, therefore, it could not be considered a cause.

AC2 is the most complex clause and is divided into two parts. AC2(a) says that for $(\vec{X} = \vec{x})$ to be a cause of Φ, there must be a setting $(\vec{X} \leftarrow \vec{x'})$ where Φ does not hold (under the contingency $\vec{W} \leftarrow \vec{w'}$). Contingencies are necessary since, for instance, Kaspar still cannot see the lion ($KS = 0$) even if we apply the intervention where the picture is still on the table ($PS \leftarrow 0$) because Kaspar's eye would have been shut anyway ($KE = 0$). We can see that $(M, \vec{u}) \models [PS \leftarrow 0]\neg(KS = 0)$ does not hold whilst $(M, \vec{u}) \models [PS \leftarrow 0, KE \leftarrow 0]\neg(KS = 0)$ does. Clearly, the contingency where $KE \leftarrow 0$ (represented by $\vec{W} \leftarrow \vec{w}$ in Definiton 4) is necessary.

AC2(b) exists to counteract some of the "permissiveness" of AC(a) by ruling out variables in the contingency as part of the actual cause. It states that the contingency $\vec{W} \leftarrow \vec{w'}$ should have no effect on Φ as long as we have the assignment $\vec{X} \leftarrow \vec{x}$. In our example, the contingency ($KE \leftarrow 0$) alone has no effect on Φ: Kaspar still cannot see the lion. The definition states that this should be true for all subsets W' of $\vec{W}$, including the empty set[3].

Finally, AC3 asserts that the identified cause is minimal. In our scenario, it prevents $(PS = 1 \wedge KE = 1)$ from being a cause, since $(PS = 1)$ suffices to satisfy AC2. Thus, AC3 also holds and we can say that, in $(M, \vec{u})$, $(PS = 1)$ is a cause of $(KS = 0)$. A similar explanation can be made to show that $(KE = 1)$ is also a cause of $(KS = 0)$.

[3] There's a slight abuse of notation since w' might not be the same size as W'.

5 Causal Explanation for Kaspar

To analyse actual causality, we need to define a causal model for Kaspar. The model comprises variables, a state space (with all possible variable valuations), and a set of equations that describe the interaction between variables. In this section, we first explain the context of Kaspar interactive games and subsequently define our model of Kaspar's interaction with its surroundings during the said games. This representation (i.e., our causal model) allows us to find the actual causes for an effect by modifying variable values and observing whether the effect persists.

5.1 Interactive games

In the experiments, children and Kaspar interact with each other during games that have been found to require explanations regarding the robot's visual perspective [6]. The goal of each game is to assess whether a child can put themselves in Kaspar's shoes; we ask if they realise whether Kaspar can or cannot see something. For instance, we cover Kaspar's eyes with a blindfold and then ask the child if Kaspar can or cannot see a picture of a lion that sits in front of the robot. The correct is answer is that Kaspar cannot see the lion. If the child answer that Kaspar can see the lion, then we'd like to automatically generate an explanation as to why this is incorrect; in that case, it is because of the blindfold.

Explanations can be generated by building a causal model of the game, defining the Φ (that states, for example, that Kaspar cannot see the chosen animal) and determining causes for it. There are four distinct games that can be played: (i) *Picture game*: Several pictures of different animals are spread around the room. Kaspar chooses an animal and asks the child to show him its picture. (ii) *Head game*: Several pictures of different animals are spread around the room. Kaspar chooses an animal and asks the child to move the robot's head in the direction of the chosen animal. (iii) *Rotate game*: Different pictures of animals are spread around a turntable. Walls are put in between animals such that Kaspar can only see one animal at a time. The robot chooses an animal and asks the child to spin the table so that the correct animal is visible. (iv) *Cube game*: A cube that contains pictures of different animal on each facet is given to the child. Kaspar chooses an animal depicted in the cube and asks the child to show the correct animal.

In all of these four games, a button can be pressed by a human which causes Kaspar to fall asleep (by closing his eyes). Furthermore, a blindfold can also be used to cover Kaspar's eyes. After the child has acted on Kaspar's instruction, a human asks the child two questions: "Is Kaspar seeing or not seeing any animal?" and "Is Kaspar seeing or not seeing the chosen animal". The first questions assesses whether the child realises that Kaspar cannot see the animal because, for example, he's asleep or he's wearing a blindfold. The second checks whether the child realises that Kaspar is not seeing the chosen animal, since for instance the picture of the animal is too far away or to the left of Kaspar's field of vision.

If the child answers the questions correctly Kaspar plays a sound of the chosen animal as a reward. Otherwise, an explanation is given to the child.

5.2 Causal model

We can cover all of the four games above with a single causal model. Table 1 shows the variables in our model. During the games where Kaspar asks the child to show a picture of a certain animal, there may be multiple pictures of different animals and even other objects placed around the room. However, in our model we are only interested in asking questions about one particular object, i.e., the chosen animal. Given the situation where the child picks up an object, our model captures whether the chosen object is the correct animal. Further, we also make note of its position relative to Kaspar's line of sight (to the left or right, above or below, too far or too close), and whether the animal is in the correct orientation (e.g., not upside down).

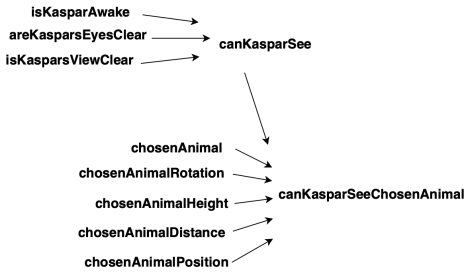


Fig. 1: Causal Network

Variables	Possible Values
chosenAnimal	correct | wrong
chosenAnimalPosition	correct | left | right
chosenAnimalRotation	correct | wrong
chosenAnimalHeight	correct | high | low
chosenAnimalDistance	correct | far | close
isKasparAwake	correct | asleep
areKasparsEyesClear	correct | covered
isKasparsViewClear	correct | obstructed
canKasparSee	true | false
canKasparSeeChosenAnimal	true | false

Table 1: Model variables

Furthermore, Kaspar's eyes can be covered by a blindfold, or Kaspar can be asleep (eyes closed), or there can be wall blocking Kaspar from seeing an animal. Any of these situations can prevent Kaspar from seeing the chosen animal. We assume that Kaspar can only see the chosen animal, if Kaspar's line of sight is clear and the animal is correctly aligned in all four senses (position, rotation, height, and distance).

A causal network is a graphical representation that displays variable dependency. In this work, we can only determine cause for acyclic models. That is, if the value of variable A depends on the value of variable B, then the opposite must not be true. Figure 1 depicts the causal network of our system. Furthermore, the structural equations that define values of the variables in the model can be seen below.

- $\mathcal{F}_{canKasparSee}() == isKasparAwake = correct \wedge$
  $areKasparsEyesClear = correct \wedge$
  $isKasparsViewClear = correct$

– $\mathcal{F}_{canKasparSeeChosenAnimal}() = canKasparSee \wedge$
$chosenAnimal = correct \wedge$
$chosenAnimalPosition = correct \wedge$
$chosenAnimalRotation = correct \wedge$
$chosenAnimalHeight = correct \wedge$
$chosenAnimalDistance = correct$

In our model, we have that the variable $canKasparSee$ is true if, and only if, $isKasparAwake$, $areKasparsEyesClear$, and $isKasparViewClear$ are all set to *correct*. The value of the variable $canKasparseeChosenAnimal$ is true if, and only if, $canKasparSee$ is true and the chosen animal is correctly positioned.

As for effects, to put it simply, in Halpern and Pearl's theory, they are represented via a Boolean combination of variable values. In this work, we are particularly interested in the effects that describe instances of Kaspar not seeing the chosen animal. This directly correlates to the two questions that can be asked to the children: "Is Kaspar seeing or not seeing any animal?" and "Is Kaspar seeing or not seeing the chosen animal". The effects we are interested in observing are (i) "Kaspar is not able to see any animal" and (ii) "Kaspar is not seeing the chosen animal". Their mathematical representation in our causal analysis syntax is: (i) $canKasparSee = false$ and (ii) $canKasparSeeChosenAnimal = false$, respectively.

Example 5 (Covered Eyes). Consider the scenario where the chosen animal is positioned correctly, however, Kaspar's eyes are covered (see Table 2). In this case, we have that our effect is $canKasparSee = false$.

We then ask the child "Is Kaspar seeing or not seeing any animal?". If the child answers the question incorrectly (in this case, by saying that Kaspar is able to see them), then we provide the explanation of why this is incorrect.

An actual cause of the type $(\vec{X} \leftarrow \vec{x})$ can only be determined if the three clauses hold. For this example, the only possible cause is ($areKasparsEyesClear$ = *covered*). AC1 and AC3 are trivially satisfied. The former, because both the cause ($areKasparsEyesClear$ = *covered*) and the effect ($canKasparSee = false$) are true in our model. The latter, because the causal explanation is minimal as there is only one variable involved. The remaining clause, AC2, is satisfied because if the value of the variable $areKasparsEyesClear$ is changed to *correct* ($areKasparsEyesClear \leftarrow correct$), we have that $\neg(canKasparSee = false)$, thus AC2. No contingency is necessary for this example.

Example 6 (Animal out of sight and Kaspar is asleep). Now, consider a second scenario. This time, Kaspar is asleep and the chosen animal is too far away. Table 3 depicts the variables and their values in this causal model. We then ask the child "Is Kaspar seeing or not seeing the lion?". The correct answer to this questions is "Kaspar is not seeing the lion", and, thus, our effect is "$canKasparseeChosenAnimal = false$". We now present the cause behind it.

Similarly to the previous example, the actual cause needs to satisfy the three clauses. This time, there are two separate causes: ($isKasparAwake$ = *asleep*)

Table 2: Example 1

Variables	Value
chosenAnimal	correct
chosenAnimalPosition	correct
chosenAnimalRotation	correct
chosenAnimalHeight	correct
chosenAnimalDistance	correct
isKasparAwake	correct
areKasparsEyesClear	covered
isKasparsViewClear	correct
canKasparSee	false
canKasparSeeChosenAnimal	false

Table 3: Example 2

Variables	Value
chosenAnimal	correct
chosenAnimalPosition	correct
chosenAnimalRotation	correct
chosenAnimalHeight	correct
chosenAnimalDistance	far
isKasparAwake	asleep
areKasparsEyesClear	correct
isKasparsViewClear	correct
canKasparSee	false
canKasparSeeChosenAnimal	false

and ($chosenAnimalDistance = far$). The explanations for both cases are similar and, thus, we'll only focus on explaining the latter case. The first clause, AC1, is satisfied because both the cause ($chosenAnimalDistance = far$) and the effect ($canKasparseeChosenAnimal = false$) are true in our model. The third clause, AC3, is also satisfied since ($chosenAnimalDistance = far$) is minimal.

Finally, AC2 is satisfied with the use of the contingencies. We apply the intervention where ($isKasparAwake \leftarrow correct$) as a contingency and we assess the parts a and b of clause AC2. The AC2(a) clause, which checks whether changes to the cause negates the effect, is satisfied since the causal formula $(M, \vec{u}) \models [chosenAnimalDistance \leftarrow correct, isKasparAwake \leftarrow correct]$ $\neg(canKasparseeChosenAnimal = false)$ holds. Similarly, AC2(b), which checks whether the contingencies applied in AC2(a) do not negate the effect, is satisfied since, for instance, $(M, \vec{u}) \models [chosenAnimalDistance \leftarrow far, isKasparAwake \leftarrow correct](canKasparseeChosenAnimal = false)$ also holds.

6 Mechanisation

We mechanise the process of causal analysis using an automated rule based system that produces a proof of causality. Our approach is advantageous over a search-based approach for causal analysis, because the latter involves building the state-space of all counterfactuals and searching through them. The process for determining causes is represented in Figure 2.

The user running the experiment selects the values of the variables in the user interface. This is done via specific key presses on a keyboard which are fed into our JAVA program (available at: https://bit.ly/ke-vs1-code). Then, a form that contains the values for the variables in the causal model is generated. The program builds a causal model, and evaluates whether the two effects hold (i.e., whether Kaspar is able to see and whether it is currently seeing the chosen animal). In case either of them hold, we determine the causes, which are then fed back into the user interface and mapped to the voice-based explanations played by Kaspar to the child. Even though multiple causes can be determined (e.g., the picture is both too far away and the wrong way around), during the

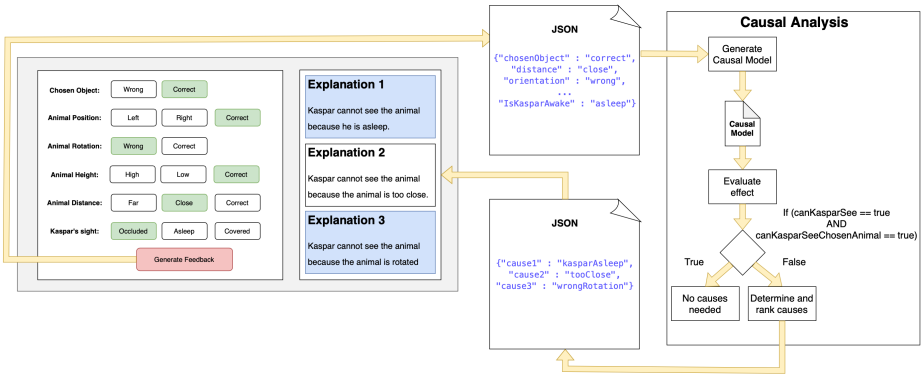


Fig. 2: Kaspar causally explains process

games, only one explanation is provided to the child to avoid overwhelming the participant. The chosen explanation is decided using an internal ranking system decided by experts; that is, in case there are multiple explanations, only the top ranked one is provided. Then, the child has a chance to correct the behaviour and the explanations are determined again, if necessary.

The rule-based system for causal analysis is proven correct for the four Kaspar games; however, it can be extend to additional games, if required. The proofs of soundness are omitted due to space limitation and are included in an extended version available online: https://bit.ly/KasparCausallyExplains

7 Explanations and their validation

We carried out an initial survey to be able to assess explanations generated by the presented system in terms of their general usefulness. For that purpose, we asked 20 adult participants (10 PhD students or staff members from research groups based at King's College London and 10 PhD students or staff members from the University of Hertfordshire) to watch videos of Kaspar providing an explanation and then rate each explanation using the explanation satisfaction (ES) scale [9]. This survey is based on several key attributes of explanations such as whether they are understandable, satisfying, sufficiently detailed, complete, informative about the interaction, useful, accurate, and trustworthy. These attributes are used to assess suitability of an explanation provided by an autonomous system. We used "what Kaspar can see" as the construct for the ES scale, which were shown to the participants for each video (cf. Table 4).

We have additionally employed the Negative Attitude towards Robots Scale (NARS) [12] to calibrate the obtained results against potential biases against robots. That allows us to later compare the current study with future studies targeting different user groups such as children. No other data has been collected and the study has been approved by the University of Hertfordshire's ethics committee for studies involving human participants, protocol number: SPECS/SF/UH/04944. Participants were provided with an information sheet

Table 4: Adapted ES scale that was shown to study participants for each video.

#	Question
1	From the explanation, I **understand** what Kaspar can see.
2	This explanation of what Kaspar can see is **satisfying**.
3	This explanation of what Kaspar can see has **sufficient detail**.
4	This explanation of what Kaspar can see seems **complete**.
5	This explanation of what Kaspar can see tells me **how to interact** with it.
6	This explanation of what Kaspar can see is **useful to my goals**.
7	This explanation of what Kaspar can see shows me how **accurate** it is.
8	This explanation lets me judge when I should **trust and not trust** Kaspar.

describing the study. Implied consent was obtained at the beginning of the survey, giving participants the option to withdraw from the study at any time.

In total, we have shown 16 videos (available at: https://bit.ly/ke-vs1-videos) to participants that contain all possible explanations for the variables of the causal network (Table 1) of the interactive games identified in [6] and described in Section 5.1. Table 5 provides an overview of the videos and describes the utterance that Kaspar uses, which is accompanied by matching gestures.

Because participant ratings were not normally distributed, we used the non-parametric one-sample Wilcoxon rank-sum test [16] to test whether ratings on the Explanation Satisfaction (ES) scale were greater than the mean value. Results attest that, when averaging across all the videos, each of the explanation is rated significantly above the neutral value (all $p < 0.001$), cf. Fig. 3a. Likewise, rating across the explanations are rated above neutral for each of the videos (all $p < 0.001$) as shown in Fig. 3b. Participant ratings on NARS attested a low

Table 5: Video recordings of explanations for the variables of the causal network that have been shown to the participants. In this table, "..." at beginning of utterances stands for "I cannot see the animal, because".

#	Variables	Game	Utterance
1	chosenAnimal: wrong	Picture	That is not the animal I have asked to see
12	chosenAnimal: wrong	Rotate	That is not the animal I have asked to see
14	chosenAnimal: wrong	Cube	That is not the animal I have asked to see
2	chosenAnimalPosition: left	Cube	...you are holding it too far left
5	chosenAnimalPosition: right	Cube	...you are holding it too far right
10	chosenAnimalPosition: left	Head	...my head is too far right
7	chosenAnimalPosition: right	Head	...my head is too far left
9	chosenAnimalDistance: far	Picture	...you are holding it too far
13	chosenAnimalDistance: close	Picture	...you are holding it too close
4	chosenAnimalHeight: low	Picture	...you are holding it too low
15	chosenAnimalHeight: high	Picture	...you are holding it too high
6	chosenAnimalRotation: wrong	Cube	...you are holding it the wrong way around
3	isKasparsViewClear: obstructed	Rotate	...the wall is in front of it
11	isKasparsViewClear: obstructed	Cube	...there is something in the way
8	areKasparsEyesClear: covered	Cube	...my eyes are covered
16	canKasparSeeAnimal: false	Cube	...you are not holding it in front of my eyes

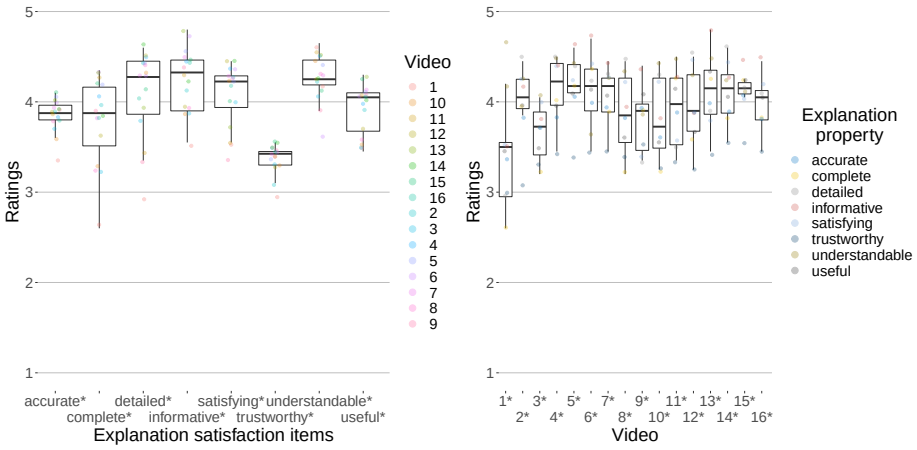
negative attitude towards robots with mean values for $\overline{S1} \approx 1.78$ (interaction subscale), $\overline{S2} \approx 2.7$ (social subscale), and $\overline{S3} \approx 1.48$ (emotion subscale). S1 and S3 are rated significantly below the neutral value (both $p < 0.001$) whereas S2 can not be reliably distinguished from neutral ($p \approx 0.053$).

These results confirm that, with healthy adults, the explanations that the system can generate are beneficial to relate cause and effect. Participants consistently rate them as accurate, complete, sufficiently detailed, satisfying, understandable, useful to their goals, and informative about the interaction. They further help to determine when to trust the robot. Knowing that adults find the generated explanations useful gives us an estimate whether the generated explanations have a potential to help autistic children in our future experiments.

8 Conclusions

We employ causal analysis as the key ingredient in providing explanations during interaction between the Kaspar robot and children. To that end, we make use of the theory of actual causation by Halpern and Pearl; outline the scenarios in which Kaspar interacts with the children; and build a causal model that covers these scenarios. We mechanised the strategy as a Java program to automatically generate causal explanations that are provided by Kaspar in order to enrich the interactions and improve trust. We validated the explanations via a controlled survey to show that they clarify and enhance the games.

For more complex interactions, we believe alternative causal explanations can be automatically ranked, e.g., in terms of their brevity. Developing theo-



(a) By Explanation property (b) By video number

Fig. 3: Results of the ES scale (5-point Likert scale) grouped by explanation property (3a) as highlighted in Table 4 and grouped by video number (3b) as listed in Table 5. Coloured points indicate the mean values of the other dimension. Asterisks mark items significantly greater than the average value.

ries of ranked explanations and empirically evaluating them in this context are worthwhile avenues of future work. Moreover, we are currently preparing extensive user studies at our partner schools to further evaluate our results for user groups involving children with ASD. Subsequently, we plan to perform iterative experiments to measure the effectiveness of explanations in improving VPT.

Acknowledgements. This work has been supported by the UKRI TAS Hub, Grant Award Reference EP/V00784X/1 and UKRI TAS Node in Verifiability, Grant Award Reference EP/V026801/2.

References

1. C. Baier, C. Dubslaff, F. Funke, S. Jantsch, R. Majumdar, J. Piribauer, and R. Ziemek. From verification to causality-based explications. In *Proc. of ICALP 2021*.
2. I. Beer, S. Ben-David, H. Chockler, A. Orni, and R. J. Trefler. Explaining counterexamples using causality. *Formal Methods Syst. Des.*, 40(1):20–40, 2012.
3. H. Chockler, O. Grumberg, and A. Yadgar. Efficient automatic STE refinement using responsibility. In *Proc. of EACAS 2008*, pp. 233–248. Springer, 2008.
4. H. Chockler and J. Y. Halpern. Responsibility and blame: A structural-model approach. *JAIR*, 22:93–115, 2004.
5. H. Chockler, D. Kroening, and Y. Sun. Explanations for occluded images. In *Proc. of ICCV 2021*, pp. 1214–1223. IEEE, 2021.
6. M. S. Gou, G. Lakatos, P. Holthaus, L. Wood, M.R. Mousavi, B. Robins, and F. Amirabdollahian. Towards understanding causality–a retrospective study of using explanations in interactions between a humanoid robot and autistic children. In *Proc. of (RO-MAN)*. IEEE, 2022.
7. J. Y. Halpern. *Actual Causality*. MIT Press, 2016.
8. A. F. C. Hamilton, R. Brindley, and U. Frith. Visual perspective taking impairment in children with autistic spectrum disorder. *Cognition*, 113(1):37–44, 2009.
9. R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*, 2018.
10. G. Lakatos, L. J. Wood, D. S. Syrdal, B. Robins, A. Zaraki, and K. Dautenhahn. Robot-mediated intervention can assist children with autism to develop visual perspective taking skills. *Journal of Behavioral Robotics*, 12(1):87–101, 2021.
11. F. Leitner-Fischer and S. Leue. Causality checking for complex system models. In *Proc. of VMCAI*, pages 248–267. Springer, 2013.
12. T. Nomura, T. Kanda, and T. Suzuki. Experimental investigation into influence of negative attitudes toward robots on human–robot interaction. *AI & Society*, 20(2):138–150, 2006.
13. B. Robins, K. Dautenhahn, R. Te Boekhorst, and A. Billard. Robotic assistants in therapy and education of children with autism. *Universal access in the information society*, 4(2):105–120, 2005.
14. M.D. Rutherford and F. Subiaul. Children with autism spectrum disorder have an exceptional explanatory drive. *Autism*, 20(6):744–753, 2016.
15. L. J. Wood, A. Zaraki, B. Robins, and K. Dautenhahn. Developing Kaspar: a Humanoid Robot for Children with Autism. *International Journal of Social Robotics*, 13(3):491–508, 2021.
16. R. F. Woolson. Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials*, pages 1–3, 2007.