



OPEN ACCESS

EDITED BY

Selma Dündar-Coecke,
University College London, United Kingdom

REVIEWED BY

Guanchao Feng,
Stony Brook University, United States
Maximilian Diehl,
Chalmers University of Technology, Sweden

*CORRESPONDENCE

Jaime Maldonado
✉ jmaldonado@uni-bremen.de

RECEIVED 22 January 2025

ACCEPTED 26 May 2025

PUBLISHED 20 June 2025

CITATION

Maldonado J, Krumme J, Zetzsche C,
Didelez V and Schill K (2025) Robot pouring:
identifying causes of spillage and selecting
alternative action parameters using
probabilistic actual causation.
Front. Cognit. 4:1565059.
doi: 10.3389/fcogn.2025.1565059

COPYRIGHT

© 2025 Maldonado, Krumme, Zetzsche,
Didelez and Schill. This is an open-access
article distributed under the terms of the
[Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in
other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication
in this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Robot pouring: identifying causes of spillage and selecting alternative action parameters using probabilistic actual causation

Jaime Maldonado^{1*}, Jonas Krumme¹, Christoph Zetzsche¹,
Vanessa Didelez² and Kerstin Schill¹

¹Cognitive Neuroinformatics, University of Bremen, Bremen, Germany, ²Department of Biometry and Data Management, Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany

In everyday life, we perform tasks (e.g., cooking or cleaning) that involve a large variety of objects and goals. When confronted with an unexpected or unwanted outcome, we take corrective actions and try again until achieving the desired result. The reasoning performed to identify a cause of the observed outcome and to select an appropriate corrective action is a crucial aspect of human reasoning for successful task execution. Central to this reasoning is the assumption that a factor is responsible for producing the observed outcome. In this paper, we investigate the use of probabilistic actual causation to determine whether a factor is the cause of an observed undesired outcome. Furthermore, we show how the actual causation probabilities can be used to find alternative actions to change the outcome. We apply the probabilistic actual causation analysis to a robot pouring task. When spillage occurs, the analysis indicates whether a task parameter is the cause and how it should be changed to avoid spillage. The analysis requires a causal graph of the task and the corresponding conditional probability distributions. To fulfill these requirements, we perform a complete causal modeling procedure (i.e., task analysis, definition of variables, determination of the causal graph structure, and estimation of conditional probability distributions) using data from a realistic simulation of the robot pouring task, covering a large combinatorial space of task parameters. Based on the results, we discuss the implications of the variables' representation and how the alternative actions suggested by the actual causation analysis would compare to the alternative solutions proposed by a human observer. The practical use of the analysis of probabilistic actual causation to select alternative action parameters is demonstrated.

KEYWORDS

robot pouring, causality, probabilistic actual causation, causal discovery, action-guiding explanations

1 Introduction

Pouring the content of a source container into a target container requires planning, perception, and action capabilities. Humans excel at these capabilities and can skillfully pour any material into containers of arbitrary shapes and dimensions. If spillage occurs, we can take corrective actions (e.g., selecting a target container with an appropriate capacity)

and try again to pour without spilling. The ability to take corrective actions is a crucial aspect of human reasoning for successful task execution. Implementing similar reasoning capabilities in robotic systems can reduce task failures and enable a robust operation in unstructured environments.

In the pouring task, possible causes for the undesired outcome of spilling the poured material include the low capacity of the target container or the diameter difference of the containers (imagine pouring from a wide glass into the narrow mouth of a bottle), among others. It is reasonable to assume that corrective actions we take in everyday life target the perceived actual cause of the undesired outcome. The extreme opposite of this behavior would consist of randomly changing action variables and observing the task outcome until finding a suitable solution. For example, if we perceive that the cause of spillage is the rim of the target container being too narrow, the corrective action would consist of selecting a target container with a wider rim. Implementing similar reasoning capabilities on robotic or automatic systems would require (1) a mechanism to identify the (not necessarily unique) actual cause of an observed outcome and (2) a principled way to determine how the causal variable needs to be changed to obtain a different outcome.

Within the context of causal analysis and modeling methods, the concept of *actual cause* refers to the conditions under which a particular event is recognized to be responsible for producing an outcome (Pearl, 2009a). Definitions of actual causation have been mainly used in practical applications to generate explanations for observed outcomes (see works reviewed in Section 3). It has been argued that explanations obtained from the analysis of actual causation can guide the search for alternative actions aiming to change the observed outcome (Beckers, 2022). The practical utility of the analysis of actual causation for the search and selection of alternative actions remains to be explored in real-life applications.

In this paper, we explore the use of actual causation analysis to select action parameters in a robot pouring task. The aim is to identify the actual cause of spillage and to determine how a task parameter should be changed to pour without spilling. We use the probabilistic actual causation definition (Fenton-Glynn, 2021) to identify the cause of spillage among the variables involved in the task. In a series of examples, we illustrate how the analysis of actual causation can be used to select alternative task parameters. Section 2 introduces the probabilistic actual causation framework and its potential use for action guidance. In Section 2.1, we propose a procedure to use the actual causation probabilities as a principled criterion to find alternative actions.

The robot pouring task was implemented in a simulation (described in Section 4.1) using a physics engine for realistic behavior. The simulation enabled us to generate trials covering a large combinatorial space of trial parameters (fullness levels and container properties), which would have been cumbersome to achieve in a physical setup. There are two prerequisites to perform an analysis of probabilistic actual causation: (1) a causal graph of the system and (2) the conditional probability distributions necessary to compute interventional queries (i.e., “do” operations). The variables used to represent the pouring task as a causal graph are described in Section 4.2. To obtain the graph structure, we used a causal discovery algorithm (Section 4.3). The conditional

distributions used to compute the do-operations were estimated using neural networks (Section 4.4). Subsequently, the causal probability and actual causation expressions that result from the graph structure are presented in Section 4.5.

Considering that causal probability computations cannot tell *per se* what caused an observed outcome as this may be specific to the given circumstances, in Section 5.1, we analyze the causal probability functions for interventions on each variable to gain insight into their influence on spillage and to emphasize their limitations for action guidance.

Subsequently, in Sections 5.2 and 5.3 we demonstrate that the analysis based on actual causation can be used in a principled way to select alternative action parameters. Firstly, in Section 5.2 we present four detailed examples of applying actual causation analysis on spillage trials to select alternative action parameters. The examples illustrate how different alternative actions yield lower or higher probabilities of spillage. Secondly, in Section 5.3 we evaluate the likelihood of finding an alternative value and the pouring success rates obtained when running the spillage trials using alternative values. The evaluation was conducted using a test dataset acquired in simulation. Overall, Sections 5.2 and 5.3 provide empirical evidence of the practical usefulness of the actual causation approach to identify alternative parameters to prevent spillage. Finally, in Section 6, we discuss the implications of the variables’ representation and how the alternative actions suggested by the actual causation analysis would compare to the alternative solutions proposed by a human observer.

In summary, our contributions are:

- We report a complete analysis of probabilistic actual causation on a practical problem with all the necessary steps for its implementation (analysis of the task, variable definition, causal-graph structure, and estimation of causal probabilities using neural networks).
- We show how to use the explanations obtained with the analysis of probabilistic actual causation for action guidance in a practical use case. By doing this, we go beyond the basic diagnostic (attribution of actual cause) toward action guidance.
- We evaluate the capability of the actual causation approach to automatically identify alternative parameters of different variables.
- We evaluate the pouring success rates obtained when the alternative parameters are used to prevent spillage.

2 Probabilistic actual causation and action-guiding explanations

The concept of actual cause refers to the conditions under which a particular event is recognized to be responsible for producing an outcome in a specific scenario or context (Pearl, 2009a). The purpose of determining an actual cause is to find which past actions explain an already observed output (Beckers, 2022). The definition of actual causation proposed by Halpern and Pearl (2005) for deterministic scenarios is one of the most prominent definitions in the literature (Borner, 2023). Fenton-Glynn (2017,

2021) proposed an extension of Halpern and Pearl’s definition that is apt for probabilistic causal scenarios. The Fenton-Glynn’s definition of actual causation is formulated in the framework of probabilistic causal models, where the causal model is a causal Bayesian network represented graphically as a directed acyclic graph (DAG) and the link between each node/variable and its direct causes is modeled probabilistically (Pearl, 2009b). In the DAG formalism, nodes correspond to variables, and directed edges (i.e., arrows) indicate causal influences. Additionally, the $do(x)$ operator represents the operation of setting the value of a variable X to $X = x$ (i.e., the variable X is instantiated to a value x), such that $P(y|do(x))$ represents the probability of obtaining the outcome Y that would result from setting X to x by means of an intervention. Fenton-Glynn (2021) proposes the following definition of actual causation¹:

Probabilistic Actual causation 1. Within a given causal model, consider a cause X of an outcome Y with a directed path \mathcal{P} from X to Y ; let the variables that are not on \mathcal{P} be denoted by \mathbf{W} , and the set of mediators on \mathcal{P} by \mathbf{Z} . Given the actually observed values $(x, \mathbf{w}^*, \mathbf{z}^*)$, we say that X taking the value $X = x$ rather than $X = x'$ is the actual cause of event Y (or $Y = 1$) when the following probability raising holds for all subsets \mathbf{Z}' of \mathbf{Z} :

$$P(Y | do(\mathbf{W} = \mathbf{w}^*, X = x, \mathbf{Z}' = \mathbf{z}^*)) > P(Y | do(\mathbf{W} = \mathbf{w}^*, X = x')). \tag{1}$$

It is important to emphasize that the term actual cause refers to token causal relations, as opposed to type causal relations (Pearl, 2009a; Halpern and Pearl, 2005; Fenton-Glynn, 2017). That is, an actual cause refers to a specific scenario, where the causal statements of the definition are regarded as *singular, single-event*, or *token-level* (Pearl, 2009a). This is evident in the inequality (1), which compares the probabilities of an event or outcome Y given the observed values of the variables in \mathbf{W} and \mathbf{Z}' , which can be seen as the given context.

In general, the notion of actual causation is considered the key to constructing explanations (Pearl, 2009a). Beckers (2022) has used the term *action-guiding explanations* for scenarios where the analysis of actual causation aims to find explanations for the outputs produced as the result of performing an action. Beckers (2022) suggests that actual causes can be used for action guidance because they enable the identification of alternative actions that provide better or worse explanations of an outcome. Beckers (2022) provides a conceptual analysis² of how action-guiding explanations relate to *sufficient* and *counterfactual* explanations, two other forms of potentially action-guiding explanations. In Beckers’ account, a sufficient explanation indicates the conditions under which an action guarantees a particular output. On the other hand, a counterfactual explanation informs which variables would have had to be different (and in what way) for the outcome to be different. Beckers (2022, p. 2) concludes that actual causes stand between

1 The notation and terminology follow the definition PC1 presented by Fenton-Glynn (2021, p. 72).

2 The conceptual analysis conducted by Beckers (2022) originally focused on notions of causal explanations in the context of deterministic causal models. Nevertheless, we consider that his conceptual accounts can be applied to probabilistic causal models.

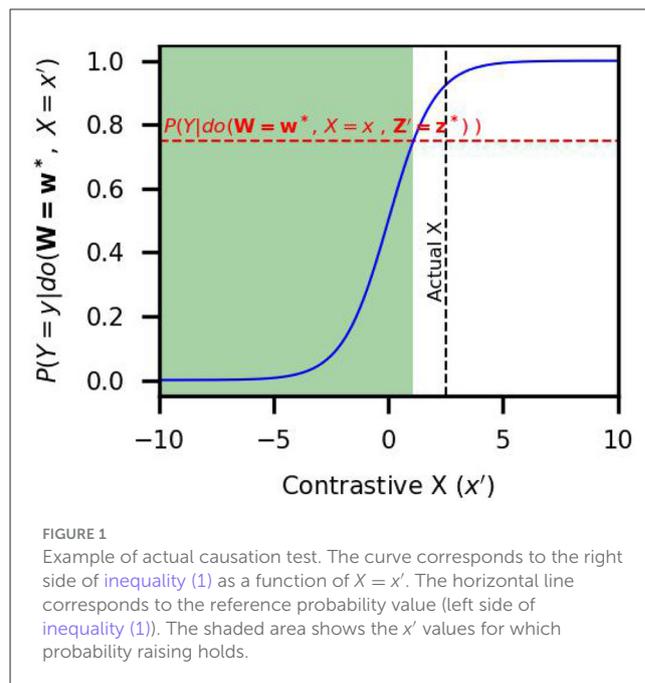


FIGURE 1
Example of actual causation test. The curve corresponds to the right side of inequality (1) as a function of $X = x'$. The horizontal line corresponds to the reference probability value (left side of inequality (1)). The shaded area shows the x' values for which probability raising holds.

sufficient and counterfactual explanations: “an actual cause is a part of a good sufficient explanation for which there exist counterfactual values that would not have made the explanation better.”

2.1 Using the actual causation inequality to select alternative actions

In inequality (1), $P(Y|do(\mathbf{W} = \mathbf{w}^*, X = x, \mathbf{Z}' = \mathbf{z}^*))$ provides a reference value to check for actual causation. This reference value takes into account the actual values of the variables. Using a contrastive value $X = x'$ on the right side of inequality (1), $P(Y|do(\mathbf{W} = \mathbf{w}^*, X = x'))$ is used to check whether or not probability raising holds, thereby providing a principled criterion to determine whether X taking the value $X = x$ rather than $X = x'$ is the actual cause of event Y .

When the variables are continuous, $P(Y|do(\mathbf{W} = \mathbf{w}^*, X = x'))$ can be plotted as a function of the contrastive value $X = x'$. This is illustrated in Figure 1. The comparison against the reference probability $P(Y|do(\mathbf{W} = \mathbf{w}^*, X = x, \mathbf{Z}' = \mathbf{z}^*))$ reveals the values x' for which probability raising holds (shaded region in Figure 1).

Fenton-Glynn’s framework of actual causation provides contrastive explanations in a given context (Borner, 2023). By definition, inequality (1) demands that the explanation remains valid when holding fixed all variables in \mathbf{W} and (the subsets) \mathbf{Z}' at their actual values; probability raising need not hold for other contexts. In this work, we propose to use the contrastive explanations obtained by applying the actual causation inequality for action guidance. Given that an outcome Y has been observed, the goal is to select an alternative action that will prevent the outcome. Recalling that probability raising entails that X taking the value $X = x$ rather than $X = x'$ is the cause of event Y , we propose to select x' values as alternative actions. In the case

illustrated in Figure 1, while selecting an X value in the shaded area much smaller than the actual $X = x$ is likely to change the outcome, selecting X close and above the actual value will likely leave the outcome unchanged. We emphasize that the magnitude of probability raising might drastically differ within the range of contrastive values. This has implications for the suitability of different x' values as alternative actions. Consider the alternative actions $X = -5$ and $X = 0$ in the schematic example illustrated in Figure 1. Based on the probabilities, it can be assumed that $X = -5$ is a better choice since selecting $X = 0$ has a higher chance of leaving the outcome unchanged. Under these considerations, the selection of an alternative parameter can be based on a pre-defined probability threshold.

An alternative parameter can be found automatically with the following steps:

1. Identify the range of contrastive values where probability raising holds using inequality (1).
2. Within these values, select the subset of values with a probability below a pre-defined probability threshold.
3. From this second subset, select the closest value to the current parameter.

In the last step we use a distance criterion. However, other application-dependent criteria can be used (e.g., the cost or availability of different alternatives) to select an alternative parameter. The crucial aspect of the automatic search is that the probability threshold is selected such that it becomes very likely to change the outcome.

In summary, to check for actual causation using inequality (1), one must decide which variable takes the role of X and, given the graph structure, identify the variables for the sets \mathbf{W} and \mathbf{Z}' . Thus, the analysis of actual causation can be applied to different variables. In the robot pouring task, if spillage occurs, the aim is to find alternative parameters to repeat the action. In this work, we use the probabilistic actual causation framework to guide the selection of alternative parameters in a principled way. In Section 5.2, we illustrate in a series of detailed examples the impact of using different probability thresholds (i.e., 0.2 for low spillage probability and 0.5 for chance-level probability) on the identified alternative parameters. Subsequently, in Section 5.3, we evaluate the pouring success rates obtained using the alternative parameters identified using a 0.1 probability threshold.

3 Related work

This section presents robotics applications that use causal methods related to our work (i.e., actual causation, causal Bayesian networks, and causal discovery). We start by presenting applications that use the Halpern-Pearl definition of actual causality (Halpern and Pearl, 2005). Subsequently, we review works that use causal discovery to learn the structure of causal Bayesian networks to model robotic tasks in different contexts. Finally, we review the method proposed by Diehl and Ramirez-Amaro (2023) to predict and prevent failures using causal-based contrastive

explanations, highlighting the similarities and differences to our approach.

Araujo et al. (2022) use the actual causation framework by Halpern and Pearl in a human-robot interaction setting where a robot interacts with children with Autism Spectrum Disorder (ASD). The robot plays different interactive games with the children, aiming to improve the children's ability to see the world from the robot's point-of-view. The authors present a tool that uses a causal model of the interactive games and the actual causation framework. This is applied to explain events during the game's course. For example, if the robot cannot see an object involved in the interaction, it explains to the child why it cannot see it (e.g., "I cannot see it because it is too high") (Araujo et al., 2022). The actual cause of an event is analyzed using a rule-based system, as opposed to a search over the possible counterfactuals (Araujo et al., 2022). The usefulness of the explanations generated by the system was evaluated by asking a group of observers to watch videos of the robot providing explanations in different situations and then rate each explanation. The rating was based on qualitative criteria, e.g., whether the explanation was understandable, sufficiently detailed, or informative about the interaction, among other aspects.

Zibaei and Borth (2024) use the Halpern and Pearl actual causation framework to retrieve explanations of failure events in unmanned aerial vehicles (UAVs). In this context, an actual causation analysis aims to provide actionable explanations, that is, an explanation that indicates which corrective actions can be taken to prevent future failures. The analysis of actual failure causes was applied to different UAV failure scenarios (loss of control, events during take-off and cruise, and equipment problems). The causal models of failure events were constructed using flight logs recorded at run-time containing abstracted events and raw sensor data. In order to diagnose instances of a particular type of failure (e.g., instances of crash events in the logs), the causal graph of the failure and the actual values of the monitored event and sensor data were analyzed using a tool for the automatic checking of the Halpern and Pearl actual causation conditions. The correctness of the diagnoses was evaluated using a manually labeled ground-truth dataset.

Chockler et al. (2021) designed an algorithm to explain the output of neural network-based image classifiers in cases where parts of the classified object are occluded. The algorithm applies concepts of the Halpern and Pearl definition of actual causation to generate explanations. The explanation consists of a subset of image pixels, which is the minimal or approximately minimal subset that allows the neural network to classify the image. It can be tested if a pixel is a cause of the classification by considering a subset of pixels in the image that does not include that pixel. In this case, applying a masking color to any combination of pixels from this subset does not alter the classification output. However, if we apply a masking color to the entire subset along with the individual pixel, this will lead to a change in the classification result. In this way, the algorithm ranks pixels according to their importance for the classification. The authors evaluate the explanations of their system by comparing their results with the outputs of other explanation tools. They compare the size of the explanation (smaller is better) and the size of the intersection of the explanation with the occluding object (smaller is better), getting favorable results for their approach. Additional work on explaining

image classifiers was done by Chockler and Halpern (2024) and has been applied to extend the work of Kommiya Mothilal et al. (2021), who used a simplified version of the actual causation framework. The authors show that extending the previous work to the full version of the framework is possible and can benefit future work in explaining image classifiers. However, they suggest that using the full definition might make computing the explanations more complex.

The reviewed works show that practical applications of the concepts of actual causation have focused on generating explanations, leaving the potential to guide decision-making processes aside. To the best of our knowledge, the probabilistic actual causation definition of Fenton-Glynn (2017, 2021) has not been used in any practical application.

Causal methods have been applied to model generic robot tasks such as pushing, pick-and-place, and stacking (Ahmed et al., 2020; Brawer et al., 2020; Huang et al., 2023; Diehl and Ramirez-Amaro, 2023) and context-specific ones such as human-robot interaction (Catri et al., 2022) and household tasks (Li et al., 2020). In the context of tool affordance learning, causal discovery has been used to identify the effect of push and pull actions on an object's final position in a real setup (Brawer et al., 2020). To reduce the sim-to-real gap in robot-object trajectories, a custom causal discovery algorithm was used to optimize the simulated physical parameters (Huang et al., 2023). Additionally, recent work has used simulations to explore large combinatoric spaces of task parameters and causal methods to model task outcomes (Ahmed et al., 2020; Huang et al., 2023; Diehl and Ramirez-Amaro, 2023).

Similar to our work, causal methods have been applied to find causal-based contrastive explanations for task failures (Diehl and Ramirez-Amaro, 2022). In subsequent work, this method has been extended to predict and prevent task failures by finding corrective parameters (Diehl and Ramirez-Amaro, 2023). Examples of contrastive explanations are provided for a cube stacking task and for dropping spheres into different containers (bowls, plates, and glasses) (Diehl and Ramirez-Amaro, 2022), and the method for prediction and prevention has been applied to a cube stacking task (stacking one cube and stacking three cubes) (Diehl and Ramirez-Amaro, 2023).

In particular, the method proposed by Diehl and Ramirez-Amaro (2023) has methodological similarities to our approach. Their method uses a causal Bayesian network to predict errors and probabilities of success to find a corrective action. Simulated data are used to learn the causal Bayesian network's structure and estimate its joint probability distribution. When an action is predicted to fail, a search is conducted in a discretized parameter space to find the parametrization that needs the least interval changes to achieve a successful execution based on the predicted success probability (Diehl and Ramirez-Amaro, 2023). For example, searching for close parametrizations in the cube-stacking task means that starting from a robot hand position that will likely fail leads to finding the closest position that will likely succeed (Diehl and Ramirez-Amaro, 2023).

The search criterion is based on contrastive explanations (Diehl and Ramirez-Amaro, 2022), which compare the variable parametrization of the failed action and the closest parametrization that exceeds a success probability threshold (ϵ) (Diehl and

Ramirez-Amaro, 2023). The success probabilities are retrieved from the factorized form of the joint probability distribution. The search is conducted within a tree that contains a complete parametrization of the parent variables of the outcome variable. Since the search is conducted within a tree structure, the time to find an alternative parametrization depends on the success probability threshold, the number of parent variables, and the number of discrete intervals of the variables (Diehl and Ramirez-Amaro, 2023).

In Table 1, we summarize the similarities and differences between the method proposed by Diehl and Ramirez-Amaro (2023) and our approach. Their method automatically finds the corrective parameters, which might involve changing the value of one or more variables. In contrast, our approach is restricted to searching for corrective parameters in a single variable. Their method identifies the closest parametrization that is predicted to succeed. On the other hand, our approach identifies a range of values that explain the outcome according to the probability-raising criterion. Within this range of values, a specific value likely to yield success can be selected using a probability criterion (see explanation in Section 2.1). Similarly to Diehl and Ramirez-Amaro (2023), a success probability threshold can be used as a selection criterion.

4 Materials and methods

4.1 Robot pouring simulation

The robot pouring task was simulated using CoppeliaSim Version 4.8.0 (rev. 0) (Rohmer et al., 2013) with the Open Dynamics Engine (ODE), a physics engine for rigid body dynamics and collision detection. The setup consists of an UR5 robot arm with a parallel jaw gripper, as shown in Figure 2. In the simulation, the robot poured marbles from a source container into a target container. The marbles were simulated with CoppeliaSim's particle object, which simulates spherical particles using parameters for their diameter (1.5 cm) and density (2,829.42 kg/m³, empirically determined).

In each trial, the source container (capacity = 514.72 cm³) was filled with simulated marbles (Figure 2a). The characteristics of the pouring movement were fixed for all trials. The robot grasped the source container and brought it to a pouring position relative to the target container's rim (Figure 2b). In the pouring position, the gripper was rotated with a fixed rotation velocity (1 rad/s) until reaching a fixed angle of -15° (Figure 2c). After pouring all the marbles (Figure 2d), spillage was detected using a force sensor located on the base of the target container. In each trial we manipulated the amount of marbles in the source container and the dimensions of the target container. The dimensions of the source container remained fixed across trials. In each trial, the source container was filled with a random amount of marbles, and a target container with random dimensions (height and rim diameter) was generated. The positioning of the marbles inside the source container was random. Six thousand pouring trials were simulated. The rationale of the random trial parameters and variable definitions are explained in Section 4.2. Details of the random distributions used to sample the trial parameters are

TABLE 1 Comparison of methods.

	Similarities					Differences		
	Model	Training data	Type of explanation	Selection of corrective action	Input	Parameter space	Corrected variables	Contrasted probabilities
Diehl and Ramirez-Amaro (2023)	Causal Bayesian network (discrete variables)	Simulation	Contrastive	Based on a probability threshold	Current values of all variables (i.e., parametrization) and probability of the outcome variable	Discrete	The corrective parametrization might involve changing one or more variables	Conditional probabilities from the factorized joint probability distribution
Ours	Causal Bayesian network (continuous and discrete variables)	Simulation	Contrastive	Based on a probability threshold	Current values of all variables (i.e., parametrization), causal probability (do-operator) of the outcome variable, and variable to correct	Continuous	The corrective parametrization changes one variable	Causal probabilities (do-operator)

provided in Table 2. In each trial, the randomness of the spillage outcome results from the behavior of the particles simulated by the physics engine and its interaction with the characteristics of the target container.

4.2 DAG variables

In each trial, the source container was filled to a random fullness level, and a target container with random dimensions was generated. The rationale for the poured amounts and the target container characteristics in each trial aims to capture typical spillage causes. Table 2 provides details of the variable definitions and random sampling distributions. The target container can have an equal, smaller, or larger capacity than the source container. The capacity of the target container influences the probability of spillage (spillage is more likely to occur when pouring into a target container of a small capacity). We represent this cause of spillage with the variable *relative capacity* (*RC*).

Aside from the target container’s capacity, spillage depends on the complex interplay between the fullness level of the source container and the rim dimensions of the source and target containers. Imagine pouring marbles from a wide glass into a bottle through its narrow mouth. While pouring a single marble is likely to succeed, the amount spilled will increase as the number of marbles increases. We represent the fullness level of the source container with the variable *fullness* (*FU*).

The target container’s rim can be equal, smaller, or larger in diameter than the source container. The relation between rim diameters influences the probability of spillage (spillage is more likely to occur when pouring into a target container with a smaller rim diameter). We represent this cause of spillage with the variable *relative diameter* (*RD*). As a result, in each trial, the dimensions of the target container (height and rim diameter) are determined by the *RC* and *RD* variables.

We also consider the fact that spillage occurs when the amount of marbles exceeds the capacity of the target container. The source containers’ capacity and fullness level determine the poured amount (recall that all the marbles are poured). We represent the difference in volume between the poured amount and the target container’s capacity with the variable *relative volume* (*RV*). Finally, we represent the outcome of the pouring trial with the variable *spillage* (*S*), a binary variable that indicates whether or not spillage occurred. *S* is labeled as *true* irrespective of the number of spilled marbles (i.e., spilling one or twenty marbles yields *S = true*).

4.3 Determination of DAG structure using causal discovery

In order to perform an analysis of actual causation, we require a DAG of the causes of spillage, where edges represent probabilistic effects between variables. A naive approach to setting the causal structure would be to assume that each variable described in Section 4.2 is connected to the outcome *S* with an edge. This would constitute a strong assumption in which all variables are direct causes of *S*, excluding the possibility of indirect effects. In this

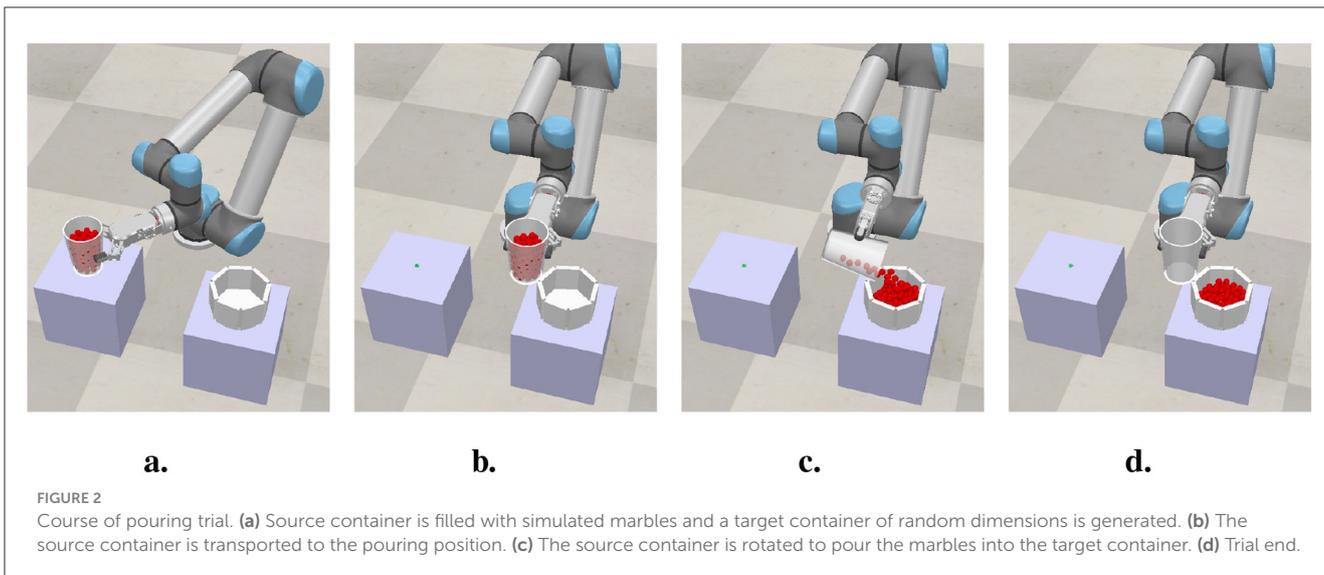


TABLE 2 Variable definitions and sampling of trial parameters.

Variable	Acronym	Definition and sampling distribution
Relative capacity	RC	RC represents the relation between the containers' capacities, defined as $RC = \frac{\text{target capacity}}{\text{source capacity}}$. Sampled from a truncated Gaussian distribution $\mathcal{N}_{trunc}(\mu = 1.0, \sigma = 0.25, \text{min} = 0.5, \text{max} = 2.0)$. $RC < 1$ corresponds to target containers of lower capacity, and $RC > 1$ to target containers of larger capacity.
Fullness	FU	FU expresses the fullness level of the source container as a fraction. Sampled from $\mathcal{N}_{trunc}(\mu = 0.7, \sigma = 0.2, \text{min} = 0.3, \text{max} = 1.0)$. For example, $FU = 0.5$ corresponds to a half-full source container.
Relative diameter	RD	RD represents the relation between the container rim diameters, defined as $RD = \frac{\text{target diameter}}{\text{source diameter}}$. Sampled from $\mathcal{N}_{trunc}(\mu = 1.0, \sigma = 0.25, \text{min} = 0.5, \text{max} = 1.5)$. $RD < 1$ corresponds to target containers of smaller diameter, and $RD > 1$ corresponds to target containers of larger diameter. The target container's height and diameter were determined based on the sampled RC and RD.
Relative volume	RV	RV represents the volume relation between the poured amount and the target container's capacity, defined as $RV = \frac{\text{poured volume}}{\text{target capacity}}$. $RV < 1$ indicates that the poured amount fits into the target container, and $RV > 1$ indicates that the poured amount exceeds the target container's capacity.
Spillage	S	Binary variable indicating whether or not spillage occurred.

respect, it is important to recall that the validity of the analysis of actual causation relies on the correctness of the DAG structure. To avoid making naive assumptions about the causal structure, we leverage the availability of simulated pouring trials to determine the structure of the DAG using a causal discovery algorithm.

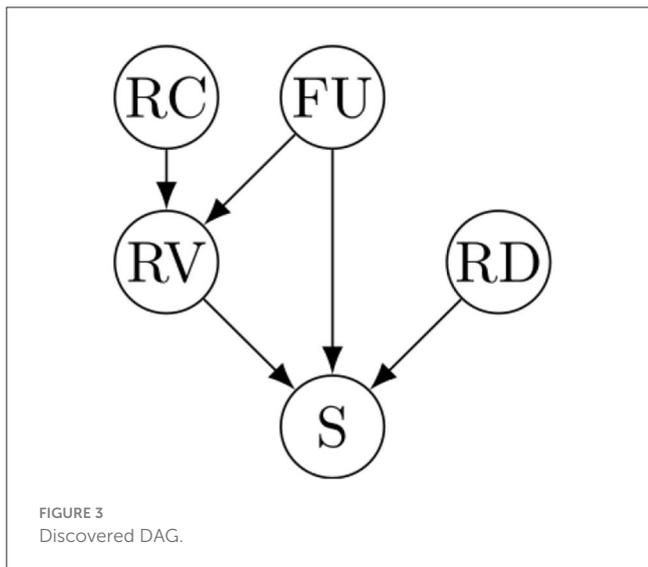
In order to determine the structure of the DAG, the variables described in Section 4.2 were processed with the PC algorithm (Spirtes and Glymour, 1991), a well-established causal discovery algorithm (Glymour et al., 2019; Nogueira et al., 2022). In general, causal discovery algorithms, also known as structure learning algorithms, perform a systematic analysis of many possible causal structures, typically by testing probabilistic independence and dependence between the variables (Malinsky and Danks, 2017; Glymour et al., 2019; Nogueira et al., 2022). We used the PC implementation available in Tetrad (version 7.6.5-0),³ a software toolbox for causal discovery (Ramsey et al., 2018). As a statistical test, we use the Degenerate Gaussian Likelihood Ratio Test (DG-LRT) (Andrews et al., 2019), which has shown

good discovery performance on datasets containing continuous and discrete variables (Andrews et al., 2019). We executed the algorithm on 1,000 bootstraps of the data to ensure the stability and reliability of the inferred causal relationships (if the results vary widely over the different bootstrap samples, the output of the algorithm is considered unstable) (Glymour et al., 2019). Further details about the PC parameters, bootstrapping results, and modeling assumptions are provided in the Supplementary material. The discovered DAG is shown in Figure 3.

4.4 Estimation of do-probabilities using neural autoregressive density estimation

In this paper, we use *neural autoregressive density estimators* (NADEs) (Garrido et al., 2021) to estimate the conditional distributions necessary to compute the interventional queries in inequality (1). NADEs are used as universal approximators and are flexible to work with models containing continuous and discrete variables (Garrido et al., 2021). Garrido et al. (2021)

³ Publicly available at: <https://www.ccd.pitt.edu/tools/>, access: 06.12.24.



demonstrated in a series of examples that NADEs provide a practical modeling architecture to estimate causal quantities from models with linear and non-linear relationships between variables of different distributions (e.g., Normal, Log-normal, and Bernoulli) using the DAG formalism and the do-calculus framework.

A DAG model represents the conditional dependencies between the set of J random variables X_1, \dots, X_J . The DAG induces a joint probability distribution $P(X)$ of the variables, which can be factorized into the conditional distributions of each X_j conditioned on a function f_j of its parents $PA(X_j)$ (the causal Markov condition):

$$P(X) = \prod_j P(X_j | f_j(PA(X_j))) \tag{2}$$

A NADE is a generative model that estimates the conditional distributions in Equation 2 (Garrido et al., 2021). The functions f_j are parametrized as independent fully connected feed-forward neural networks. Therefore, there is a neural network for each variable in the DAG. Each neural network takes the parents of the variable of interest $PA(X_j)$ and outputs the parameters of the distribution of X_j . For example, the neural network outputs the mean and standard deviation of Gaussian variables. These networks are trained using the negative log-likelihood of Equation 2 as the loss function. The individual conditional distributions in Equation 2, also called independent causal mechanisms or Markov Kernels, are used to estimate the effects of interventions (Garrido et al., 2021). The causal estimates are reliable under the assumption that the DAG structure is correct and that the training data provides enough support to learn the distribution parameters (Garrido et al., 2021). Details of the implementation are provided in the Supplementary material.

4.5 Causal probability expressions and actual causation inequalities

The factorized joint probability distribution that results from the DAG structure shown in Figure 3 is expressed as a product of conditional distributions and independent causal mechanisms:

$$P(RC, FU, RV, RD, S) = P(RC) \cdot P(FU) \cdot P(RV|RC, FU) \cdot P(RD) \cdot P(S|FU, RD, RV) \tag{3}$$

These conditional distributions and independent causal mechanisms are approximated using NADEs, as described in Section 4.4. The implemented neural networks are shown in Figure 4. The distribution of the continuous variables RC , FU , RV , and RD is approximated by Gaussian distributions. For these variables, each network takes the parent of the corresponding variable as input and produces two parameters, one for the mean (μ) and one for the standard deviation (σ) of the Gaussian distribution. Following the implementation of Garrido et al. (2021), the root nodes (RC , FU , and RD) take a constant value as input, represented as a “1” in Figure 4. The discrete variable S follows a Bernoulli distribution. Its neural network takes the variables RV , FU , and RD as input and outputs the probability of sampling $S = true$ given the input values. The NADE estimators are used to compute causal (or interventional) probabilities (e.g., $P(S|do(RD))$) and actual causation queries in the form of inequality (1).

4.5.1 Causal probability expressions

The DAG indicates that S has direct and indirect causes. The path $RC \rightarrow RV \rightarrow S$ shows that RC is an indirect cause of S , and its effect is mediated by RV . The causal relation between FU and S combines a direct path ($FU \rightarrow S$) and indirect path ($FU \rightarrow RV \rightarrow S$). Finally, RD is a direct cause of S (path $RD \rightarrow S$). Ultimately, the probability of S depends on the interactions between RV , FU , and RD . Additional insight into the factors that yield spillage can be gained by estimating the effects of each variable on S . Causal effects are functions of interventional probabilities $P(S|do(X))$ over different values of X ; the expressions can be obtained using the rules of do-calculus with the *causaleffect* R package (version 1.3.15) (Tikka and Karvanen, 2017). Following the procedure done by Garrido et al. (2021), we use Monte Carlo integration in order to obtain a numerical approximation of the causal expressions of interest.

Under the given causal model, the interventional probability of spillage $P(S|do(RC))$ under a given choice of RC is obtained as:

$$P(S | do(RC)) = \int_{RD, FU, RV} P(S|FU, RD, RV) \cdot P(RV|RC, FU) \cdot P(RD) \cdot P(FU) dRV dRD dFU \tag{4}$$

We approximated $P(S|do(RC))$ using Monte Carlo integration by sampling from the distributions $P(RD)$, $P(FU)$, and $P(RV|RC, FU)$, and propagating forward through the neural network implemented for $P(S|FU, RD, RV)$.

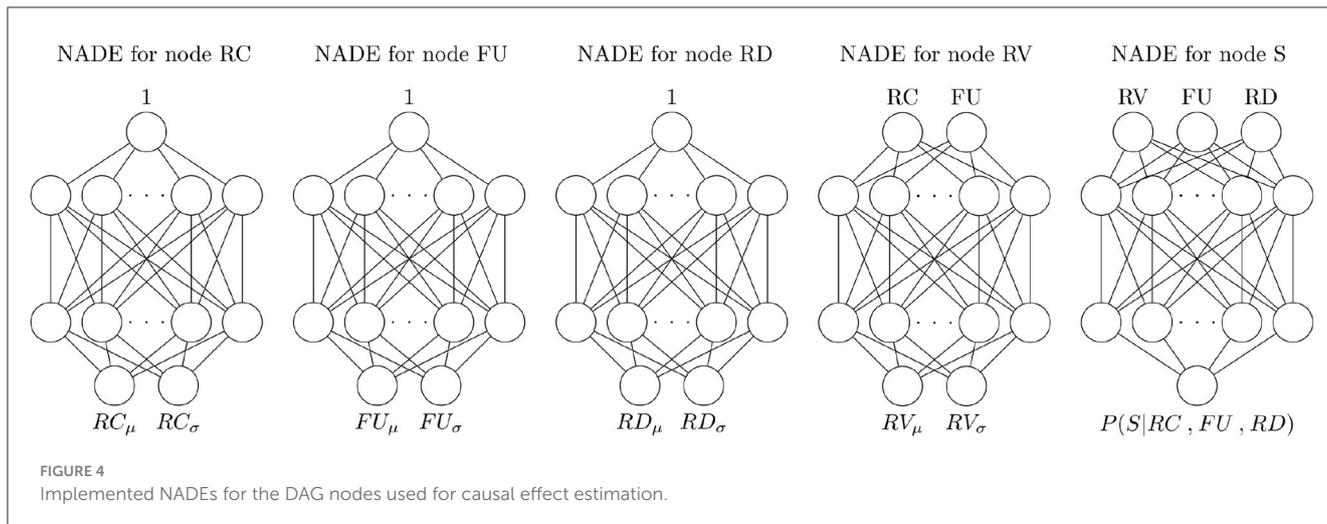


FIGURE 4 Implemented NADEs for the DAG nodes used for causal effect estimation.

The interventional probability $P(S|do(FU))$ for a given choice of FU is obtained as:

$$P(S | do(FU)) = \int_{RC, RV, RD} P(S|FU, RD, RV) \cdot P(RV|RC, FU) \cdot P(RD) \cdot P(RC) dRC dRD dRV \quad (5)$$

Similarly, $P(S|do(FU))$ was approximated using Monte Carlo integration by sampling from the distributions $P(RC)$, $P(RD)$, and $P(RV|RC, FU)$, and propagating forward through the neural network implemented for $P(S|FU, RD, RV)$.

The expression $P(S|do(FU))$ in Equation 5 combines the effect mediated by RV and the direct effect of FU . In order to assess the direct effect of FU on S we must fix RV . This enables us to assess how the direct effect of FU may differ for different choices of RV . The direct effect can be expressed by varying the values of FU with constant RV in $P(S|do(FU, RV))$, which is obtained as:

$$P(S | do(FU, RV)) = \int_{RD} P(S|RD, FU, RV) \cdot P(RD) dRD \quad (6)$$

The probability $P(S|do(FU, RV))$ was also approximated using Monte Carlo integration by sampling from the distributions $P(RD)$ and propagating forward through the neural network implemented for $P(S|FU, RD, RV)$.

Furthermore, $P(S|do(RV))$ is obtained as:

$$P(S | do(RV)) = \int_{RD, FU} P(S|RD, FU, RV) \cdot P(RD) \cdot P(FU) dFU dRD \quad (7)$$

The probability $P(S|do(RV))$ was approximated using Monte Carlo integration by sampling from the distributions $P(RD)$ and $P(FU)$, and propagating forward through the neural network implemented for $P(S|RD, FU, RV)$.

Finally, $P(S|do(RD))$ is obtained as:

$$P(S | do(RD)) = \int_{RC, FU, RV} P(S|FU, RD, RV) \cdot P(RV|RC, FU) \cdot P(FU) \cdot P(RC) dRC dFU dRV \quad (8)$$

The probability $P(S|do(RD))$ was approximated using Monte Carlo integration by sampling from the distributions $P(FU)$, $P(RC)$, and $P(RV|RC, FU)$, and propagating forward through the neural network implemented for $P(S|FU, RD, RV)$.

4.5.2 Inequalities for analysis of probabilistic actual causation

For the analysis of actual causation we identify the sets of variables necessary to compute the probabilities compared in inequality (1). In order to analyze whether RD is an actual cause of S , we consider the only path $\mathcal{P}: RD \rightarrow S$. The set of variables \mathbf{W} that lie off the path is $\mathbf{W} = \{RC, FU, RV\}$, and the set \mathbf{Z} of variables that lie intermediate between RD and S is $\mathbf{Z} = \emptyset$. RD taking the value $RD = rd$ rather than $RD = rd'$ (in the context of the observed values rc, fu, rv) is an actual cause of S (or $S = true$) when the following probability raising holds:

$$P(Y | do(\mathbf{W} = \mathbf{w}^*, X = x, \mathbf{Z}' = \mathbf{z}^*)) > P(Y | do(\mathbf{W} = \mathbf{w}^*, X = x')) \\ P(S | do(rc, fu, rv, rd)) > P(S | do(rc, fu, rv, rd')) \\ P(S | fu, rv, rd) > P(S | fu, rv, rd') \quad (9)$$

The conditional probabilities in inequality (9) were approximated using the neural network for $P(S|RV, FU, RD)$.

In order to analyze whether FU is an actual cause of S , we consider the path $\mathcal{P}: FU \rightarrow RV \rightarrow S$. The set of variables \mathbf{W} that lie off the path is $\mathbf{W} = \{RC, RD\}$, and the set \mathbf{Z} of variables that lie intermediate between FU and S on \mathcal{P} is $\mathbf{Z} = \{RV\}$. Recalling that inequality (1) is tested for all the subsets of \mathbf{Z} which includes the empty set (Fenton-Glynn, 2021), FU taking the value $FU = fu$ rather than $FU = fu'$ is an actual cause of S (or $S = true$) when the probability raising holds both for $\mathbf{Z}' = RV$ (inequality (10)) and

$Z' = \emptyset$ (inequality (11)):

$$P(Y | do(\mathbf{W} = \mathbf{w}^*, X = x, Z' = \mathbf{z}^*)) > P(Y | do(\mathbf{W} = \mathbf{w}^*, X = x'))$$

$$P(S | do(rc, rd, fu, rv)) > P(S | do(rc, rd, fu'))$$

$$P(S | rd, fu, rv) > \int_{RV} P(S | rd, fu', RV) \cdot P(RV | rc, fu') dRV \quad (10)$$

$$P(Y | do(\mathbf{W} = \mathbf{w}^*, X = x, Z' = \mathbf{z}^*))$$

$$> P(Y | do(\mathbf{W} = \mathbf{w}^*, X = x'))$$

$$P(S | do(rc, rd, fu)) > P(S | do(rc, rd, fu'))$$

$$\int_{RV} P(S | rd, fu, RV) \cdot P(RV | rc, fu) dRV > \int_{RV} P(S | rd, fu', RV) \cdot P(RV | rc, fu') dRV \quad (11)$$

The integrals in inequality (11) and on the right side of inequality (10) were approximated using Monte Carlo integration by sampling from the distribution $P(RV | RC, FU)$ and propagating forward through the neural network implemented for $P(S | FU, RD, RV)$.

5 Results

5.1 Causal probabilities

The causal probability $P(S|do(RC))$ was computed using Equation 4 over a range of RC values covering low ($RC < 1$) and large ($RC > 1$) target container capacities. The estimated probabilities are shown in Figure 5a. Over the whole range of RC, the probability lies slightly below 0.5. This probability curve reflects the fact that RV mediates the effect of RC, which in turn crucially depends on FU. Without knowing the context of FU, the total effect of RC on S lies around the chance level. Thus, solely knowing or setting a RC value does not provide enough information on causing or preventing spillage.

The probabilities $P(S|do(FU))$ and $P(S|do(FU, RV))$ were computed using Equations 5, 6, respectively, over a range of FU values covering low ($FU < 0.4$), medium ($0.4 \leq FU \leq 0.6$), and high ($FU > 0.6$) fullness levels of the source container. The estimated probabilities for $P(S|do(FU))$, shown in Figure 5b, indicate that low fullness levels yield a low probability (≈ 0.2) of spillage. This probability increases as the fullness level increases, reaching a maximum of 0.6. For the direct effect $P(S|do(FU, RV))$, we present the probabilities obtained for three levels of RV: $RV = 0.25$ (the poured amount fits into the target container), $RV = 1$ (the poured amount equals the target container's capacity), and $RV = 1.5$ (the poured amount exceeds the target container's capacity). The estimated probabilities are shown in Figure 5c. It can be observed that the probabilities obtained for $RV = 0.25$ and $RV = 1$ are similar. Thus, for $RV < 1$, increasing FU will have a similar effect on the probability of spillage. In contrast, the results obtained for $RV = 1.5$ show different non-linear behavior with overall spillage probabilities above the chance level.

The probability $P(S|do(RV))$ was computed using Equation 7 over a range of RV covering values where the poured amount fits into the target container ($RV < 1$) and the poured amount exceeds the target container's capacity ($RV > 1$). The estimated probabilities are shown in Figure 5d. For $RV < 1$, we obtained spillage probabilities ≈ 0.4 . Therefore, without knowing the context of the other variables, when $RV < 1$, the probability of spillage lies slightly below the chance level. For $RV > 1$ we observe a moderate probability increase, reaching a maximum of 0.88. By definition, $RV > 1$ indicates that the poured amount exceeds the target container's capacity. Thus, in practice, a large probability of spillage (≈ 1) for $RV \gg 1$, and a step probability increase for $RV \approx 1$ would have been expected. This suggests that the NADE for RV is smoothing the estimated probabilities.

The probability $P(S|do(RD))$ was computed using Equation 8 over a range of RD values covering target containers of smaller diameter ($RD < 1$) and target containers of larger diameter ($RD > 1$). The estimated probabilities are shown in Figure 5e. For $RD < 0.8$, we observe a constant large probability of spillage (≈ 0.9). For $0.8 \leq RD \leq 1.1$, we observe a sharp probability decrease. For $RD > 1.1$ we observe a low probability of spillage.

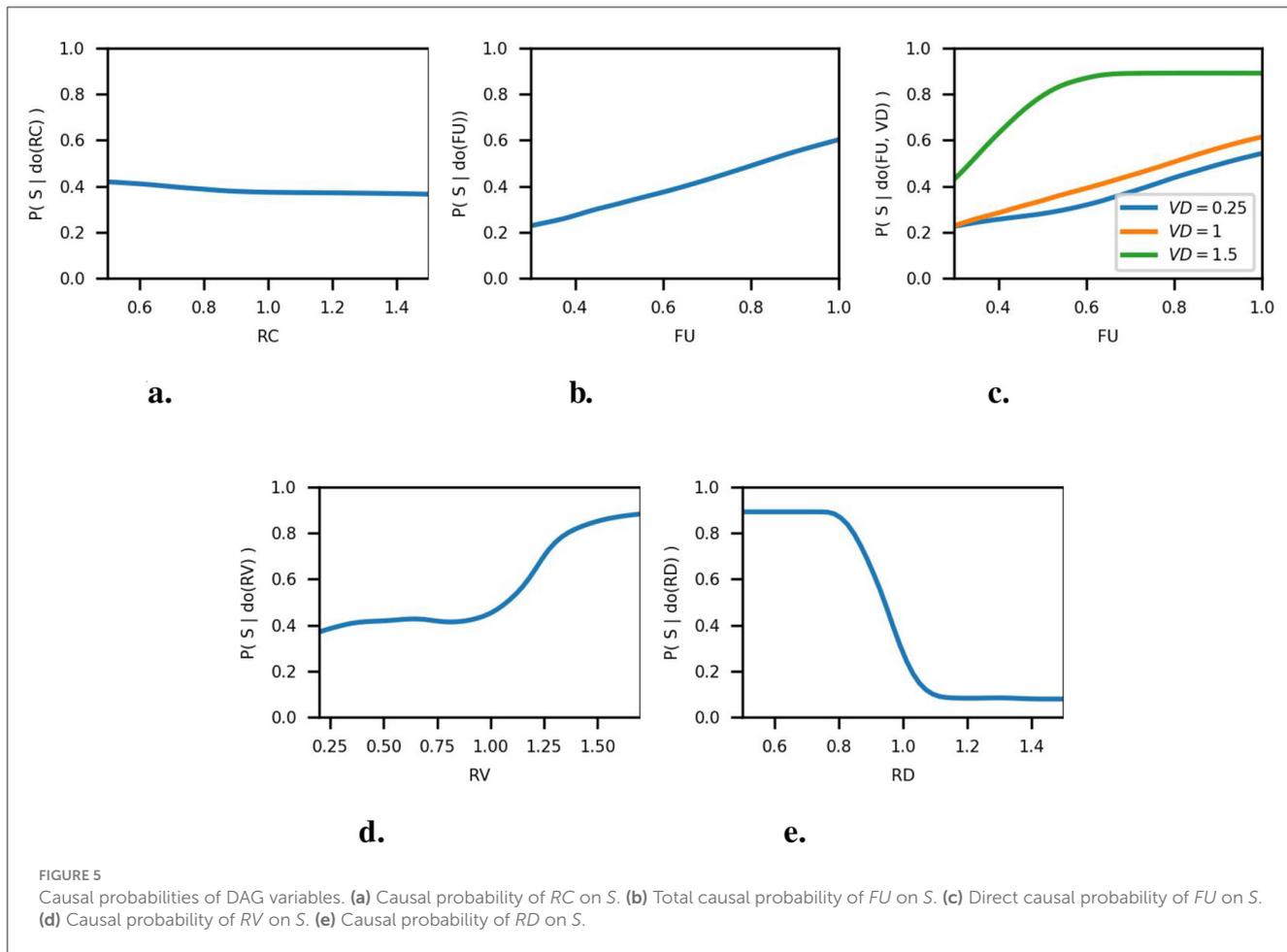
In contrast to the results obtained for $P(S|do(FU))$ and $P(S|do(RV))$, $P(S|do(RD))$ shows distinct ranges where RD yields low and high probability of spillage. Additionally, the range of values where $P(S|do(RD))$ shows a sharp probability decrease indicates that a slight change in RD can significantly affect the probability of spillage.

Overall, the causal effects presented in this section exhibit non-linear behavior. Due to the non-linear behavior and the different magnitudes of the spillage probability, using the individual causal probabilities to analyze a trial with particular RC, FU, RV, and RD values does not provide conclusive information. This happens because causal probabilities in the form of $P(S|do(X))$ do not consider the context (i.e., the actual value) of the other variables. In the following section, we present a set of examples where the actual causation framework is used to determine which variable (at its actual value) is an actual cause of spillage and how the variable should be changed to yield a different outcome.

5.2 Actual causes of spillage and alternative actions

As explained in Section 2, the actual causation inequality compares a reference probability value against the probabilities obtained from contrastive values to check whether or not probability raising holds. As proposed in Section 2.1, we use the contrastive values where probability raising holds to guide the selection of alternative parameters to avoid spillage. In this section, we present the results obtained by using the automatically-selected alternative values to achieve either a chance-level or a low probability of spillage. In the following, we explain the rationale behind the selected spillage trial examples.

According to the considerations presented in Section 4.2 and the DAG discovered from the training data (Figure 3), spillage has three direct causes: RV, FU, and RD. At the beginning of the



pouring action, i.e., at the pouring onset, the interaction of *FU* and *RD* determines the probability of spillage. Afterward, as the content is poured into the target container, spillage will occur after the poured amount exceeds the container’s capacity; that is, spillage caused by *RV*. Recalling that *RV* is caused by *FU* and *RC*, the capacity of the target will be exceeded only if $RC < 1$. The smaller the *RC*, the smaller the *FU* will be necessary to exceed the target container’s capacity. Otherwise, when $RC > 1$, spillage is caused only by *FU* and *RD*. Therefore, we exemplify the usage of the actual causation framework for the analysis of spillage trials with $0.5 < RC < 1$. For simplicity, we focus the analysis on *RD* and *FU*, which have a direct effect on *S*. We do not consider *RC* because it doesn’t have a direct effect on *S* (the effect of *RC* is mediated by *RV*, which also depends on *FU*).

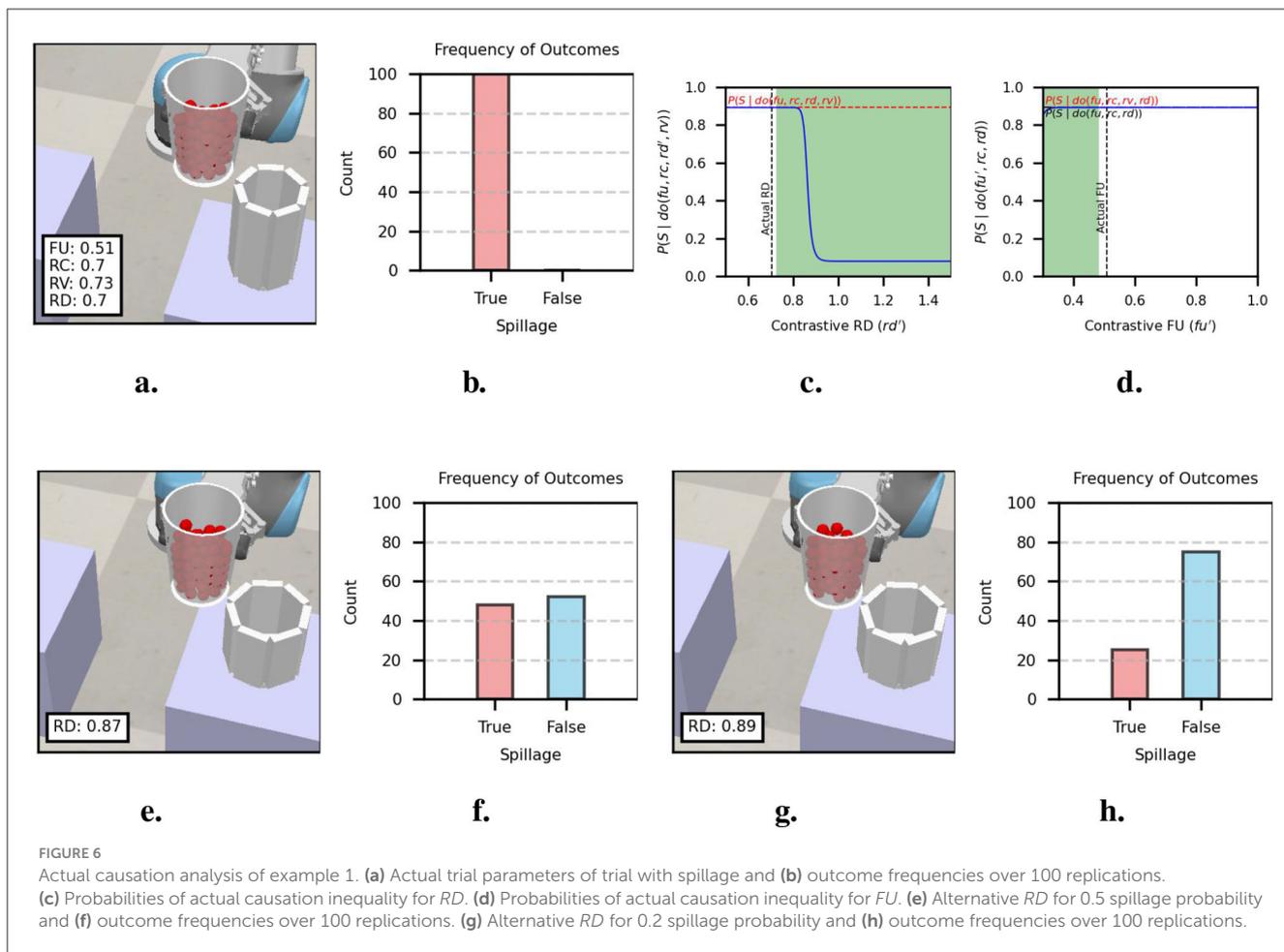
In each example, we start by presenting the frequency of the outcomes (spillage true or false) over 100 replications using the actual trial parameters. Afterward, we perform the actual causation analysis of the *FU* and *RD* variables. To compare the suitability of different alternative parameters, we select alternative *FU* or *RD* parameters where $P(Y|do(W = w^*, X = x'))$ yields a low probability (0.2) and a chance-level (0.5) probability. The alternative parameters are identified automatically following the steps described in Section 2.1. We conduct 100 replications using the alternative parameters and present the frequency of the outcomes.

5.2.1 Example 1

In this example, the source container was filled to a medium level ($FU = 0.51$). The particles were poured into a target container of smaller capacity ($RC = 0.70$) and smaller diameter ($RD = 0.70$). The actual trial parameters are shown in Figure 6a. The outcomes over 100 replications with the actual parameters, shown in Figure 6b, indicate a high probability of spillage.

Figure 6c shows the reference probability and the probabilities obtained for contrastive *RD* values obtained from inequality (9). In the area where probability raising holds, the contrastive probabilities show a sharp decrease around $RD \approx 0.8$ and a low probability for $RD > 0.9$. We select the alternative values $RD = 0.87$ (chance-level probability) and $RD = 0.89$ (low probability ≈ 0.2), as shown in Figures 6e, g, respectively. It is important to note that changing *RV* and keeping *RC* constant produce a target container of smaller height. The results from 100 replications are shown in Figures 6f, h. The replications with the alternative parameters show that $RD = 0.89$ is better suited to avoid spillage. It is also interesting to note that due to the non-linearity of the probabilities, small changes in *RD* yield a large effect on the probability of spillage.

Figure 6d shows the reference probabilities and the probabilities obtained for contrastive *FU* values obtained from Equations 10, 11. The probability curve that results from the contrastive values is nearly horizontal and very close in magnitude



to the reference probabilities. The area for which probability raising holds (i.e., the shaded region) thus results from small magnitude differences barely noticeable in Figure 6d. The contrastive values for which probability raising holds yield a large probability of spillage. Therefore, selecting an alternative FU is unlikely to change the outcome.

5.2.2 Example 2

In this example, the source container was filled to a medium-high level ($FU = 0.64$). The particles were poured into a target container of a slightly smaller capacity ($RC = 0.96$) and slightly larger diameter ($RD = 1.10$). The actual trial parameters are shown in Figure 7a. The outcomes over 100 replications with the actual parameters, shown in Figure 7b, indicate a low probability of spillage.

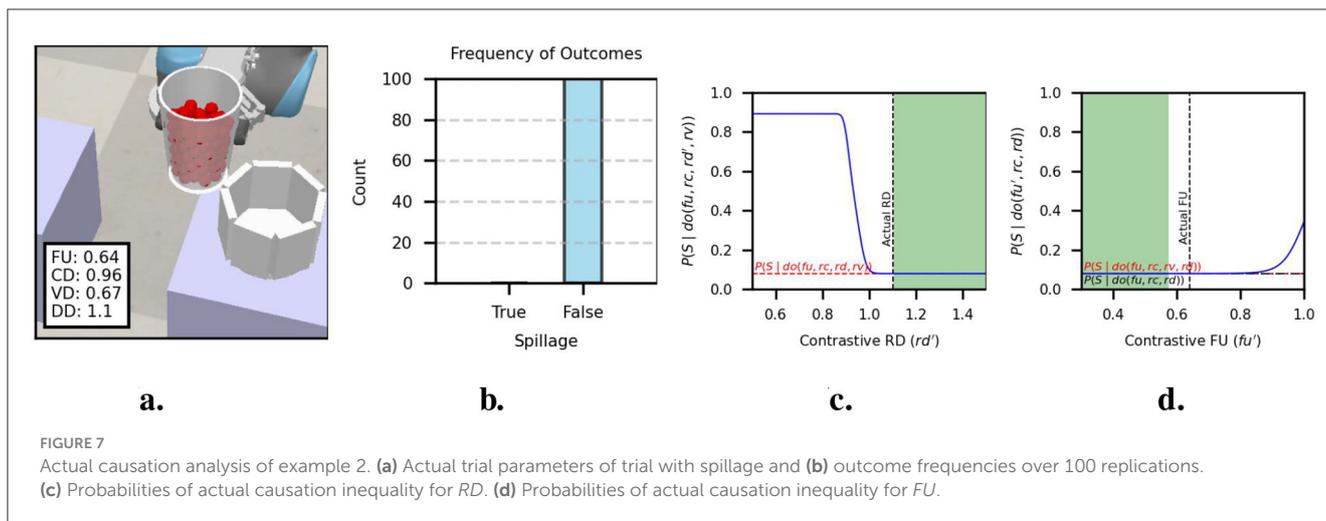
Figure 7c shows the reference probability and the probabilities obtained for contrastive RD values obtained from inequality (9). It can be observed that the contrastive values for which probability raising holds yield a low probability of spillage. Therefore, selecting an alternative RD is unlikely to change the outcome. A similar result is observed from the actual causation analysis of FU. Figure 7d indicates that selecting an alternative FU will unlikely change the outcome. In this example, the analysis of actual causation indicates that repeating the execution of the pouring action with the actual parameters is likely to succeed.

5.2.3 Example 3

In this example, the source container was filled to a high level ($FU = 0.77$). The particles were poured into a target container of smaller capacity ($RC = 0.79$) and slightly smaller diameter ($RD = 0.98$). The actual trial parameters are shown in Figure 8a. The outcomes over 100 replications with the actual parameters, shown in Figure 8b, indicate a moderately larger probability of spillage (≈ 0.6).

Figure 8c shows the reference probability and the probabilities obtained for contrastive RD values obtained from inequality (9). In the area where probability raising holds, the contrastive probabilities show a sharp decrease around $RD \approx 1.0$ and a low probability for $RD > 1.1$. We select the alternative values $RD = 0.99$ (chance-level probability) and $RD = 1.02$ (low probability ≈ 0.2), as shown in Figures 8e, g, respectively. The results from 100 replications are shown in Figures 8f, h. The replications with the alternative parameters show that $RD = 1.02$ is better suited to avoid spillage.

Figure 8d shows the reference probabilities and the probabilities obtained for contrastive FU values obtained from inequalities (10) and (11). In the area where probability raising holds, we observe low probabilities for $FU < 0.6$. For $FU > 0.6$, we observe smooth probability increments. We select the alternative values $FU = 0.75$ (chance-level probability) and $FU = 0.64$ (low probability ≈ 0.2), as shown in Figures 8i, k, respectively. The results from 100 replications are shown in Figures 8j, l. The replications with the



alternative parameters show that $FU = 0.64$ is better suited to avoid spillage.

5.2.4 Example 4

In this example, the source container was filled to a high level ($FU = 0.91$). The particles were poured into a target container of smaller capacity ($RC = 0.69$) and slightly larger diameter ($RD = 1.08$). The actual trial parameters are shown in Figure 9a. The outcomes over 100 replications with the actual parameters, shown in Figure 9b, indicate a large probability of spillage.

Figure 9c shows the reference probability and the probabilities obtained for contrastive *RD* values obtained from inequality (9). It can be observed that the contrastive values for which probability raising holds yield a high probability of spillage. Therefore, selecting an alternative *RD* is unlikely to change the outcome.

Figure 9d shows the reference probabilities and the probabilities obtained for contrastive *FU* values obtained from inequalities (10) and (11). In the area where probability raising holds, we observe low probabilities for $FU < 0.4$. For $FU > 0.4$, we observe smooth probability increments. We select the alternative values $FU = 0.76$ (chance-level probability) and $FU = 0.53$ (low probability ≈ 0.2), as shown in Figures 9e, g, respectively. The results from 100 replications are shown in Figures 9f, h. The replications with the alternative parameters show that $FU = 0.53$ is better suited to avoid spillage.

5.3 Evaluation of alternative actions to prevent spillage

In this section, we evaluate the capabilities of the analysis of probabilistic actual causation to guide the selection of alternative parameters to prevent spillage. For this evaluation, we generated a test dataset of 3,000 pouring trials. The trial parameters of the test dataset were sampled from the same distributions used for the training dataset, as described in Table 2. Spillage occurred in 1216 trials and 1784 trials were successful.

Following steps described in Section 2.1, we conducted the analysis of probabilistic actual causation on the spillage trials. We identified the range of contrastive values where probability raising

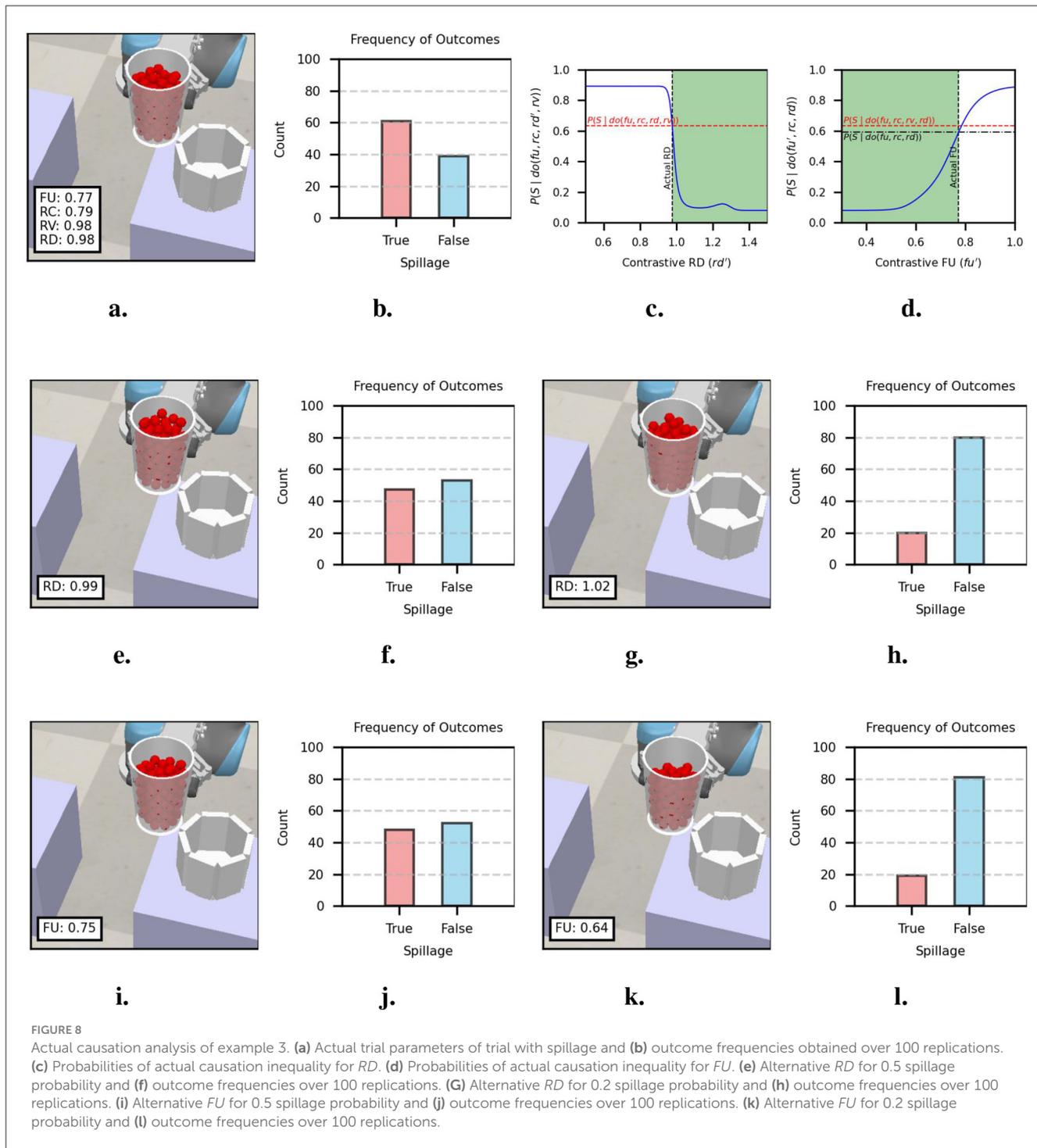
holds, and within these values, we selected the subset of values with probability of spillage < 0.1 . From this subset, the closest value to the current parameter was selected as alternative parameter. We ran the trial using the alternative parameter and recorded the outcome.

As explained in Section 2.1, the analysis of actual causation can be applied to different variables, one at a time. When inequality (1) does not hold for any of the contrastive values, the variable cannot be regarded as an actual cause. This indicates that there are no alternative values for the analyzed variable. Even if inequality (1) holds, it can also occur that the probabilities within the range of contrastive values where probability raising holds lay above the chance level or above the desired probability threshold (0.1 in our case). Therefore, it may happen that no alternative values for the variable being analyzed can be identified. To evaluate this aspect, we conducted the analysis of probabilistic actual causation on the *RC*, *FU*, and *RD* variables, and determined the percentage of the spillage trials for which an alternative parameter satisfying the probability threshold could be identified. For *RC*, an alternative parameter could be identified in 2.7% of the spillage trials, for *FU* in 59% of the spillage trials, and for *RD* in 97.9% of the spillage trials.

The marked differences between variables can be attributed to the causal structure of the task and the magnitude of the causal probabilities presented in Section 5.1. *RC* has an indirect effect on *S* through *RV*. *FU* also has an indirect effect on *S* through *RV* and a direct effect. Comparing the causal probabilities $P(S|do(RC))$ (Figure 5a), with $P(S|do(FU))$ and $P(S|do(FU, RV))$ (Figures 5b, c, respectively), it can be observed that the effect of *FU* on *S* is stronger than the effect of *RC*. Under these considerations, finding an alternative *RC* value is less likely than finding an alternative *FU* value. In contrast to *RC* and *FU*, *RD* has a direct effect on *S*, and the $P(S|do(RD))$ (Figure 5e) shows a range with low spillage probability, which leads to a higher likelihood of finding an alternative value.

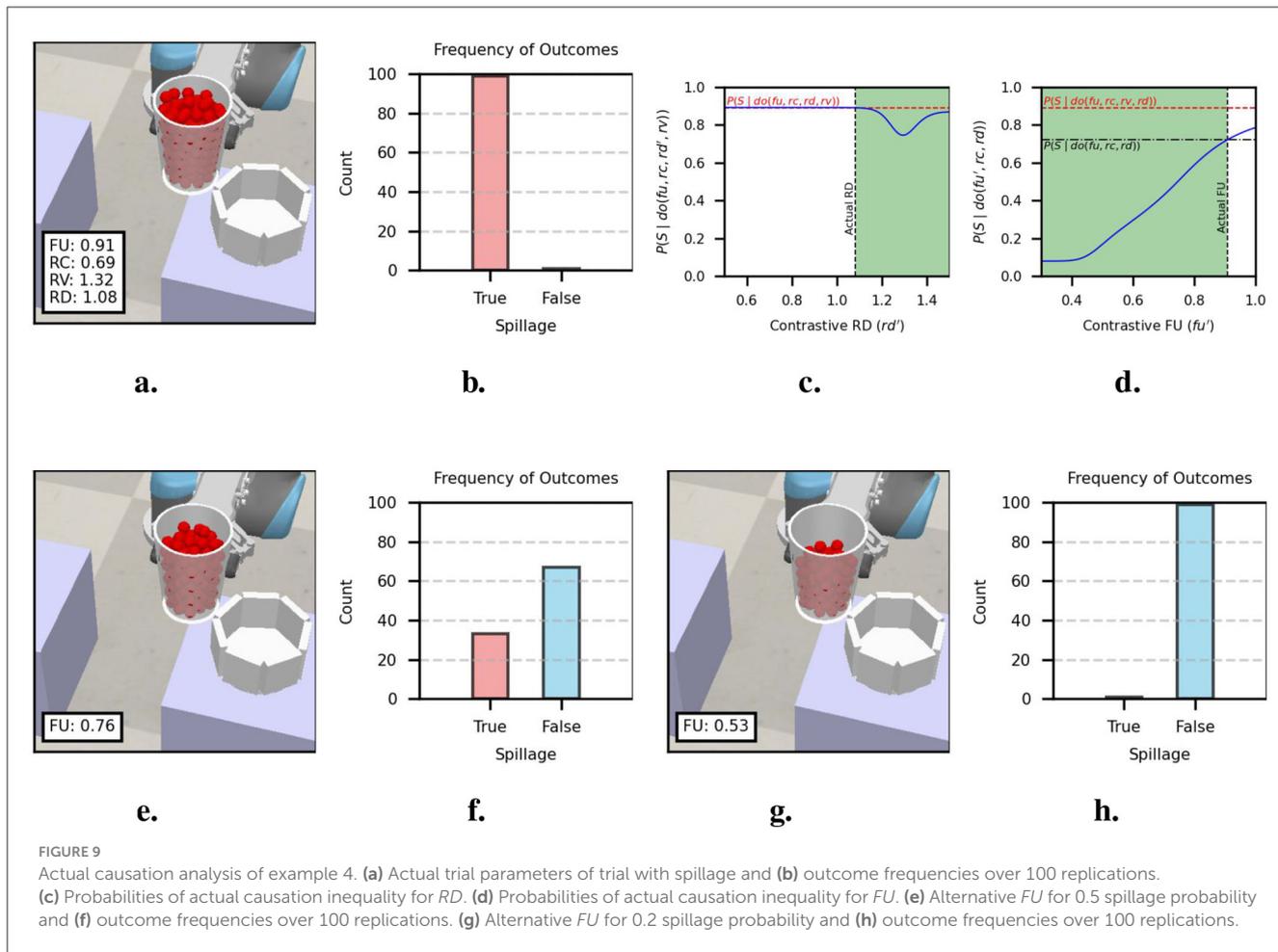
Next, we evaluate the pouring success rates obtained by running the spillage trials using the alternative *RD* and *FU* values, while keeping the other variables unchanged.⁴ As explained at the beginning of this section, in each spillage trial we set *RD* or *FU* to

⁴ *RC* was not considered for this evaluation due to the low number of trials for which an alternative value was found.



an alternative value (closest to the current value) predicted to have a spillage probability < 0.1 according to the actual causation analysis. A trial was successful if all the particles were poured into the target container. From 1,216 spillage trials, an alternative *RD* value was identified for 1,191 trials. Running these trials with the alternative *DD* values produced a success rate of 88.7%. From 1,216 spillage trials, an alternative *FU* value was identified for 718 trials. Running these trials with the alternative *FU* values success rate of 86.9%. The success rates demonstrate the practical value of the actual causation approach in identifying alternative parameters.

Finally, we evaluate the empirical success rates observed when conducting trial replications using the alternative *FU* and *RD* values. For this, we selected a random subset of 100 spillage trials from the test dataset. We ran 10 replications of each trial of the subset using the alternative *RD* or *FU* values. As a result, we obtained an empirical success rate for each trial ($\frac{\text{number of successful replications}}{10}$). The histogram of the empirical success rates obtained with the alternative *RD* values is shown in Figure 10a, and the histogram obtained for the alternative *FU* values is shown in Figure 10b. In general, the observed success



rates are in line with the probability criterion used to select the alternative values. Nevertheless, lower success rates of 0.6 and 0.7 were also obtained for some trials. Crucially, all the success rates are above the chance level, which provides empirical support for the usefulness of the alternative values.

6 Discussion

In Section 5.2, an analysis of actual causation was conducted to determine the actual cause of spillage in a set of selected spillage trials. The analysis of the FU and RD variables was conducted individually. Based on the analysis, alternative FU and RD values were selected. In the examples, we also observed cases in which the analysis indicated that no alternative FU or RD values would significantly reduce the probability of spillage.

The actual causation probabilities exhibit a non-linear behavior. In particular, a sharp decrease in the spillage probability can be observed in some ranges of RD. Considering the sharpness in the transition from low to high probability of spillage observed in the causal probability $P(S|do(RD))$ (Figure 5e) and in the actual causation probabilities in examples 1, 2, and 3 (Figures 6c, 7c, 8c, respectively), we verified that the sharp transition is not an artifact of the binary representation of the outcome. For this verification,

we examined the number of particles spilled and the percentage of spilled particles as a function of RD in the range with sharp probability transitions. Within this range of values, small changes in RD yield significant changes in the number of spilled particles and the spilled percentage. As a result of the sharp probability decrease, when the analysis indicates that a RD is the actual cause of spillage, small changes in its value significantly impact the probability of spillage (cf. examples 1 and 3). The actual causation probabilities of FU also exhibit non-linear behavior, though the ranges where the probability of spillage transitions from low to high show a less pronounced slope (cf. example 4). Nevertheless, there are cases where a small change in FU leads to significant differences in the probability of spillage (cf. example 3).

The examples show that the probabilities of actual causation provide a principled criterion for comparing alternative actions with respect to the probability of the desired outcome. In the pouring task, small differences between values separate a “bad” from a “good” alternative action due to the non-linear effect of the variables on the outcome. Based on this consideration, it is reasonable to assume that the alternative actions identified using the automatic application of the actual causation analysis might differ from those chosen by a human observer. Consider the target container dimensions from example 3 compared in Figure 11. A human observer might fail to

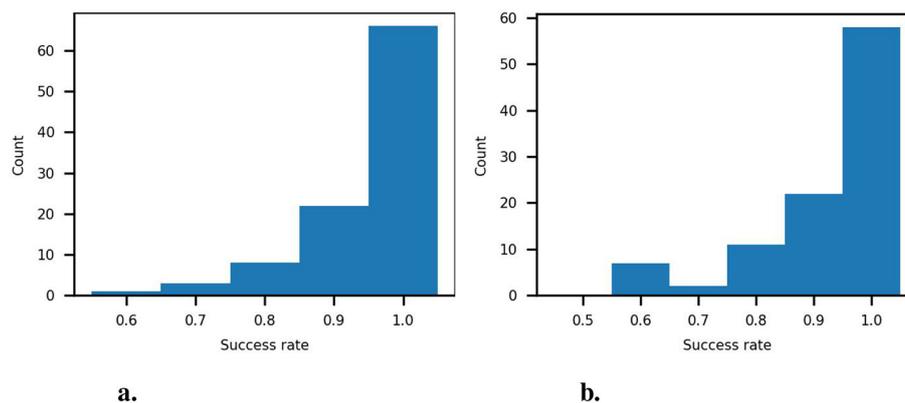


FIGURE 10 Empirical success rates obtained over 10 replications. (a) Empirical success rates using alternative *RD* values. (b) Empirical success rates using alternative *FU* values.

realize that a slight change in diameter can significantly reduce the chances of spillage due to non-linear interaction between the task variables. Therefore, it can be assumed that a human observer will choose alternative actions based on larger parameter differences than the ones suggested by the analysis of actual causation. For a human observer, a significantly larger diameter difference or a lower fullness level might result from an implicit safety margin in the selection of an alternative action to avoid spillage. Overall, human reasoning about alternative solutions will hardly resemble the analysis of actual causation, as it relies on probabilistic reasoning about the (non-linear) interaction between variables.

Regarding the usage of actual causation for action guidance, it is important to consider the availability or feasibility of alternative actions in the context of application. In simulation, generating target containers of different dimensions or changing the source container's fullness levels is straightforward. However, the available alternative actions might be limited in the real world. For example, if target containers are available only in two diameter sizes, the selection of an alternative action is reduced to making a forced choice, leaving aside any reasoning about the effect of the diameters in a continuous space on the probability of spillage. Nevertheless, even if the alternatives are limited, an analysis of actual causation can provide the agent performing or monitoring the task with useful information about possible alternative actions.

It is important to recall that the results obtained from applying the analysis of actual causation depend on the structure of the causal graph and the variable representation. In this work, we opted to represent spillage as a binary variable. However, other representations of the outcome are possible. For example, spillage could be represented as the number of spilled marbles or as a relation, such as $S = \frac{\text{number of spilled marbles}}{\text{number of marbles in the source container}}$. Defining spillage as a binary variable treats spilling one or many marbles equally. In this sense, the binary representation loses information regarding the severity of spillage. Ideally, the perfect realization of the task entails pouring without spillage. However, whether information about the severity of spillage is necessary depends on the context. For example, while spilling a few snacks

at a party would not be a problem, spilling a single particle in a chemistry laboratory might be inadmissible. The previous situational examples emphasize that the context of the application must be considered when determining the definition of the outcome variable(s).

The representation of the variables has implications for the perceptual capabilities of the agent performing or monitoring the task, be it a human or a robot. The perception of the outcome and the container properties relies on sensory cues, which might consist of visual and force feedback. For example, to determine the number of spilled particles, the agent must be able to perceive and count individual particles. The extent to which this is feasible depends on the context (e.g., counting spilled candies might be way easier than counting spilled rice grains, both for a human or a robot). In this respect, representing spillage as a binary variable has the advantage of being easier to determine, both for a human and a robot, as it requires less perception, reasoning, and action capabilities. Overall, the successful implementation of action guidance based on the analysis of probabilistic actual causation in a real-world application relies on the availability of the information necessary to compute the probabilities in [inequality \(1\)](#). For the pouring task, the agent would need an accurate perception of both containers' dimensions, the source container's fullness level, and whether or not spillage occurred.

7 Conclusions

In this paper, we conducted a probabilistic actual causation analysis of a robot pouring task. The modeling based on causal graphs and the estimation of conditional probability distributions using neural networks facilitated a qualitative and quantitative understanding of the influence of various factors on the task's outcome. Throughout a series of examples, we demonstrated that the analysis of actual causation provides a principled approach to check whether a variable is a cause of the outcome and to select alternative actions appropriate to change the observed outcome. Our results show that the analysis of actual causation

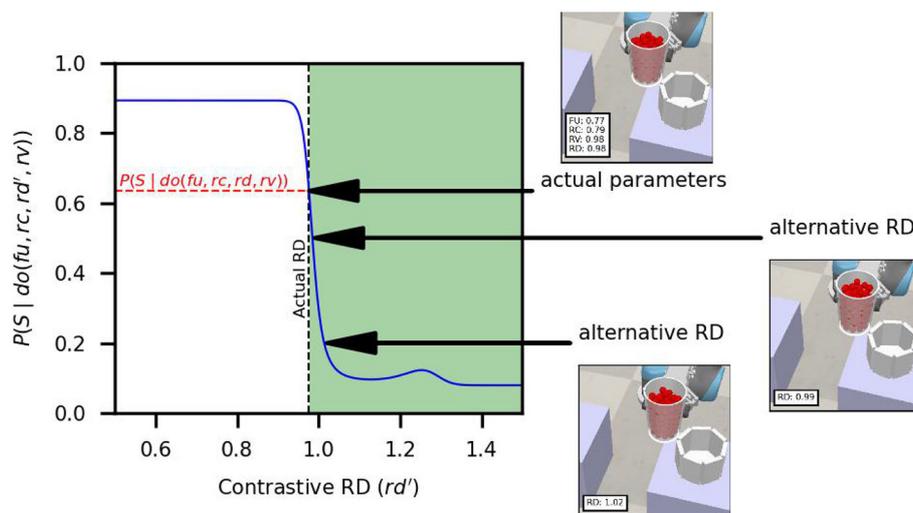


FIGURE 11
Actual causation probabilities, actual trial parameters and target containers with different RD values.

provides information about the extent to which a variable caused an observed outcome that cannot be retrieved directly from simple causal probabilities. This occurs because the causal probability of a variable on the outcome lacks information about the context of the other variables. In contrast, the definition of probabilistic actual causation considers the context of the variables and their role in the causal structure (i.e., whether the variables are mediators or are outside the causal path). In the pouring task, the analysis of actual causation enabled us to determine whether a variable was a cause of spillage and, based on the assessment of the probabilities of actual causation, the selection of an alternative action parameter.

The reliability of the analysis of actual causation relies on the correctness of the causal graph structure and the estimated conditional probability distributions. This constitutes a major challenge for implementing an actual causation analysis in real-life tasks as it requires (1) a careful selection of the variables' representation, (2) determining the structure of the causal graph, and (3) estimating conditional probability distributions. The methods described in Section 4 constitute state-of-the-art best common practices to obtain reliable and robust causal modeling results. Specifically, we used a realistic simulation of the pouring task to cover an ample combinatorial space of task parameters, which would have been cumbersome to replicate in a real environment. The simulation provided us with a large dataset to learn the causal structure of the task using a causal discovery algorithm with bootstrapping and to estimate its causal probability distributions using neural networks. In addition to the information provided in [Supplementary Section 4](#), we discuss further modeling assumptions and we provide empirical support to the correctness of the causal model.

We demonstrated the practical use of probabilistic actual causation in a robotic task. In addition, the information retrieved from actual causation analysis can be used in the context of human-machine interaction to support human decision-making. Recalling that the actual causation probabilities can be interpreted as the extent to which an alternative action parameter is a "good" or "bad"

corrective action, the framework can provide the human operator with additional information and contextual cues, for example, in augmented or virtual reality applications, to support the selection of action parameters.

In an additional scope of application, the analysis of actual causation can provide an objective baseline to evaluate the human perception of actual causes and the selection of alternative actions when a different outcome is sought. For example, the extent to which the causes perceived by a human observer correspond to the actual causes identified by the probabilistic framework can be investigated. Furthermore, the framework can be used to assess the extent to which human ratings of corrective actions in a continuum from "bad" to "good" correspond to the interpretation of the goodness of an alternative action parameter based on the actual causation probabilities used in this paper.

Data availability statement

The raw data supporting the conclusions of this manuscript will be made available by the authors, without undue reservation, to any qualified researcher.

Author contributions

JM: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. JK: Data curation, Methodology, Software, Writing – original draft, Writing – review & editing. CZ: Conceptualization, Funding acquisition, Methodology, Writing – original draft, Writing – review & editing. VD: Conceptualization, Formal analysis, Funding acquisition, Writing – original draft, Writing – review & editing. KS: Funding acquisition, Project administration, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. The research reported in this paper has been supported by the German Research Foundation DFG, as part of Collaborative Research Center (Sonderforschungsbereich) 1320 Project-ID 329551904 EASE - Everyday Activity Science and Engineering, University of Bremen (<http://www.ease-crc.org/>). The research was conducted in subproject H01 Sensorimotor and Causal Human Activity Models for Cognitive Architectures.

Acknowledgments

The implementation of the neural autoregressive density estimators was possible thanks the collaborative work and support of Konrad Gadzicki (EASE - subproject H03 Discriminative and Generative Human Activity Models for Cognitive Architectures).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships

References

- Ahmed, O., Träuble, F., Goyal, A., Neitz, A., Bengio, Y., Schölkopf, B., et al. (2020). Causalworld: a robotic manipulation benchmark for causal structure and transfer learning. *arXiv [Preprint]*. arXiv:2010.04296. doi: 10.48550/arXiv:2010.04296
- Andrews, B., Ramsey, J., and Cooper, G. F. (2019). "Learning high-dimensional directed acyclic graphs with mixed data-types," in *Proceedings of Machine Learning Research, volume 104* (Anchorage, AK: PMLR), 4–21.
- Araujo, H., Holthaus, P., Gou, M. S., Lakatos, G., Galizia, G., Wood, L., et al. (2022). "Kaspar causally explains," in *Social Robotics*, eds. F. Cavallo, J.-J. Cabibihan, L. Fiorini, A. Sorrentino, H. He, and X. Liu (Cham: Springer Nature Switzerland), 85–99. doi: 10.1007/978-3-031-24670-8_9
- Beckers, S. (2022). "Causal explanations and XAI," in *Proceedings of the First Conference on Causal Learning and Reasoning, Volume 177 of Proceedings of Machine Learning Research*, eds. B. Schölkopf, C. Uhler, and K. Zhang (Eureka, CA: PMLR), 90–109.
- Borner, J. (2023). *Causal Explanations - How to Generate, Identify, and Evaluate Them* (PhD thesis), Munich.
- Brawer, J., Qin, M., and Scassellati, B. (2020). "A causal approach to tool affordance learning," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Las Vegas, NV: IEEE), 8394–8399. doi: 10.1109/IROS45743.2020.9341262
- Castri, L., Mghames, S., Hanheide, M., and Bellotto, N. (2022). "Causal discovery of dynamic models for predicting human spatial interactions," in *International Conference on Social Robotics* (Cham: Springer), 154–164. doi: 10.1007/978-3-031-24667-8_14
- Chockler, H., and Halpern, J. Y. (2024). *Explaining Image Classifiers*. Cham: ACM. doi: 10.24963/kr.2024/25
- Chockler, H., Kroening, D., and Sun, Y. (2021). "Explanations for occluded images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (Montreal, QC: IEEE), 1234–1243. doi: 10.1109/ICCV48922.2021.00127
- Diehl, M., and Ramirez-Amaro, K. (2022). Why did i fail? A causal-based method to find explanations for robot failures. *IEEE Robot. Autom. Lett.* 7, 8925–8932. doi: 10.1109/LRA.2022.3188889
- Diehl, M., and Ramirez-Amaro, K. (2023). A causal-based approach to explain, predict and prevent failures in robotic tasks. *Robot. Auton. Syst.* 162:104376. doi: 10.1016/j.robot.2023.104376
- Fenton-Glynn, L. (2017). A proposed probabilistic extension of the Halpern and Pearl definition of 'actual cause'. *Br. J. Philos. Sci.* 68, 1061–1124. doi: 10.1093/bjps/axv056
- Fenton-Glynn, L. (2021). *Causation*. Cambridge: Cambridge University Press. doi: 10.1017/9781108588300
- Garrido, S., Borysov, S., Rich, J., and Pereira, F. (2021). Estimating causal effects with the neural autoregressive density estimator. *J. Causal Inference* 9, 211–228. doi: 10.1515/jci-2020-0007
- Glymour, C., Zhang, K., and Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Front. Genet.* 10:524. doi: 10.3389/fgene.2019.00524
- Halpern, J. Y., and Pearl, J. (2005). Causes and explanations: a structural-model approach. part I: causes. *Br. J. Philos. Sci.* 56, 843–887. doi: 10.1093/bjps/axi147
- Huang, P., Zhang, X., Cao, Z., Liu, S., Xu, M., Ding, W., et al. (2023). "What went wrong? Closing the sim-to-real gap via differentiable causal discovery," in *Proceedings of The 7th Conference on Robot Learning, Volume 229 of Proceedings of Machine Learning Research*, eds. J. Tan, M. Toussaint, K. Darvish (Atlanta, GA: PMLR), 734–760.
- Kommiya Mothilal, R., Mahajan, D., Tan, C., and Sharma, A. (2021). "Towards unifying feature attribution and counterfactual explanations: different means to the same end," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, AIES '21* (New York, NY: ACM), 652–663. doi: 10.1145/3461702.3462597
- Li, Y., Zhang, D., Yin, F., and Zhang, Y. (2020). Operation mode decision of indoor cleaning robot based on causal reasoning and attribute learning. *IEEE Access* 8:173376. doi: 10.1109/ACCESS.2020.3003343
- Malinsky, D., and Danks, D. (2017). Causal discovery algorithms: a practical guide. *Philos. Compass* 13:e12470. doi: 10.1111/phc3.12470
- Nogueira, A. R., Pugnana, A., Ruggieri, S., Pedreschi, D., and Gama, J. (2022). Methods and tools for causal discovery and causal inference. *WIREs Data Min. Knowl. Discov.* 12:1449. doi: 10.1002/widm.1449
- Pearl, J. (Ed.) (2009a). "The actual cause," in *Causality* (Cambridge: Cambridge University Press), 309–330.

that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fcogn.2025.1565059/full#supplementary-material>

- Pearl, J. (Ed.) (2009b). "Causal diagrams and the identification of causal effects," in *Causality* (Cambridge: Cambridge University Press), 65–106. doi: 10.1017/CBO9780511803161.005
- Ramsey, J. D., Zhang, K., Glymour, M., Romero, R. S., Huang, B., Ebert-Uphoff, I., et al. (2018). "Tetrad—a toolbox for causal discovery," in *8th International Workshop on Climate Informatics*. Boulder, CO.
- Rohmer, E., Singh, S. P. N., and Freese, M. (2013). "V-rep: a versatile and scalable robot simulation framework," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems* (Tokyo: IEEE). doi: 10.1109/IROS.2013.6696520
- Spirtes, P., and Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Soc. Sci. Comput. Rev.* 9, 62–72. doi: 10.1177/089443939100900106
- Tikka, S., and Karvanen, J. (2017). Identifying causal effects with the R package causaleffect. *J. Stat. Softw.* 76, 1–30. doi: 10.18637/jss.v076.i12
- Zibaei, E., and Borth, R. (2024). Building causal models for finding actual causes of unmanned aerial vehicle failures. *Front. Robot. AI* 11:1123762. doi: 10.3389/frobt.2024.1123762