How We Identify Cause-Effect Relationships Given Evidence: An Exploratory Study

Gregory Chernov

Abstract

For well-informed decision-making and precise predictions, the ability to discern causality is imperative. An innate understanding of cause-and-effect relationships in daily life can be deceptive and lead to incorrect interpretations of correlations between variables. In literature estimates of the human ability to identify causal relationships from factual information remain rare. This paper builds upon Kendall and Charles's (2022), study of the influence of persuasive false narratives on subjects' inference within a linear cause-effect schema. Drawing on Oprea's research in (2020), we extend Kendall and Charles (2022) to non-linear patterns and the relationship between accuracy and complexity. In our framework participants face repeated tasks with data generation processes (DGP) of varying structure with 3 variables (Eberhardt 2017) that they need to deduce by observing the data. Our findings reveal that even in the simplest case with three variables, the concept of conditional independence poses a significant challenge in correctly identifying cause-and-effect relationships, with accuracy levels only marginally exceeding random guessing (10-25 percent improvement). Moreover, different DGPs exhibit varied accuracy that does not align with standard t-complexity (Oprea 2020), where more connections imply greater complexity. Accuracy in our task is rather determined by the type of data source of the variable (observed or intervening) and how much the relationship or lack thereof matches the rest of the relationships in the DGP.

Keywords: complexity, structural learning, causal discovery, bounded rationality, behavioral economics, economics experiments, imperfect information decision making

JEL: C91, D91 G0, K4

1. Introduction

Understanding causality is crucial for making accurate predictions and informed decisions. While humans are adept at identifying cause-and-effect relationships in everyday situations, relying solely on observed correlations between variables in contextual judgments can be flawed. The prevailing solution to navigate this complexity is rooted in the assumption that evidence from experimental interventions, such as Randomized Controlled Trials (RCTs) or A/B tests, offers unambiguous and reliable guidance for the ascription of a causal link. The underlying principle appears straightforward: if an intervention yields an effect, a causal link exists; if not, there is no link. However, this approach oversimplifies situations where causal mechanisms exhibit a more intricate structure. Consider a scenario where a medication influences health status by altering hormone levels (a mediation effect), and these hormone levels are also independently influenced by another factor, such as sleep quality. Focusing solely on the direct impact of the intervention on taking the drug may lead to misconceptions about the total effect, especially when the combination of factors results in reversing the single effects in the opposite direction. In such cases, the observed effect might even contradict the actual influence of the medication in conjunction with sleeping.

The repercussions of misinterpreting causal relationships have garnered significant attention in economic theory, as exemplified by studies by Spiegler (2020). However, empirical investigations into the specifics and origins of these misinterpretations predominantly reside within the field of psychology. Noteworthy works, including those by Bramley et al. (2017); Schulz and Gopnik (2004), are mostly related to active learning. These studies have focused on the learning process itself, with a notable emphasis on how individuals acquire knowledge through observing outcomes and engaging in direct interaction with a process. The emphasis on firsthand experience is more relevant for studying general regularities in human learning rather than for analyzing specific aspects of decision-making. In the latter, subjects deal with accumulated information represented as data, including numerous events or statistics. The closest design (Taylor and Ahn 2012) dedicated to assessing the ability to cope with causality with data includes only observational data types (see section 2.2 for a comprehensive overview). In this study, we establish a framework where participants are presented with datasets (observational and interventional) and tasked with fully reconstructing the original process behind the data. The obtained results are of interest to both economists and cognitive scientists since the framework involves tasking participants to interact with datasets rather than single events.

To illustrate the problem statement, let us consider a scenario where someone is deciding whether or not to go on a diet. The subject observes several data columns from an arbitrary study, representing dietary adherence and health status. In this context, the decision maker might conclude a direct correlation between those factors, even though there might be others, such as exercise, which are linked with dietary adherence. Now, let us imagine that this decision-maker gains access to Dutch famine statistics from 1944-1945, a period when starvation was caused by factors unrelated to the behavior of the subjects. Would this person reconsider their conclusions once intervention data, such as the famine statistics, becomes available?

This formulation appears deceptively simple, implying practical relevance, but its simulation in a modelling environment quickly becomes intricate. In our task, we require participants to reverse engineer a causal mechanism from data generated by a predefined process. To maintain the clarity of this task, we restrict it to binary events, involve only three variables, and use data from which the original Data Generating Process (DGP) can be fully identified by observational and interventional datasets. The intervening variable in the intervention dataset is always set to one. If such a problem were to be tackled by a professional or an ideal decision-maker, they could recognize the DGP in a few mouse clicks within our task. Importantly, our task lacks special contextual information; we use the metaphor of plants influencing each other and explain the intervention by the physical manipulation of seeds, integrating it in a natural way only in the instructions. Subsequently, participants work solely with variables labeled as X, Y, and Z.

From a cognitive standpoint, engaging with causality involves a spectrum of activities, including hypothesis generation, testing, and revision, all underpinned by an understanding of the distinctions between passive observing (seeing) and tracking consequence of active manipulation (doing) (Galles and Pearl 1998). That is why, unlike tasks that involve mere correlation inference in linear setting (Kendall and Charles 2022) or learning environment (Schulz and Gopnik 2004), We consider the causal discovery to be a basis for modelling framework (Eberhardt 2007), since it provides a foundation for constructing tasks with a mix of data types—interventional and observational. The availability of data resulting from intervention manipulation presents individuals with a task akin to real-world situations. If they understand that interventions have a one-way effect and recognize that the observed correlation suggests an effect without a specific direction (though not the absence of direction), it would indicate a comprehensive understanding of the causal setting. Conversely, a lack of ability to interpret the consequences of interventions underscores fundamental misconceptions regarding causal mechanism mapping into the evidence. Given the compound nature of this challenge, our primary research interest is framed as follows: *When individuals encounter evidence from various data types, including observational, experimental, or a combination of both, which specific combinations of these data sources are most prone to leading to erroneous causal conclusions, and what factors contribute to the susceptibility to biases in such scenarios?* We employ mixed datasets because some Data Generating Processes (DGPs) require more considered information at the same time to be fully identified than others. If this applies to a perfectly rational decision-maker, the cognitive process of Homo economicus needs to address it as well, considering cognitive limitations. Therefore, we compare the recognition of different DGPs with different data sets, as detailed in Section 2.3. Asking participants to identify only one DGP at a time might be less parsimonious, but collecting more could lead to interference.

To explore this, we have developed a new framework that allows us to address these questions at both the between-subject (in treatment and control comparisons preventing from potential interference) and within-subject levels, considering different approaches provided for acquaintance in section 2.3. Given that our design incorporates within-subject comparisons, we are analyzing the structural complexity as a factor, exploiting the conceptual framework proposed by Oprea (2020) and mapping it into the context of causal discovery Eberhardt (2007). In our experimental setup, participants are tasked with observing data from three variables and deducing the underlying Data Generating Process (DGP) along with all its dependencies. This task is repeated 18 times for each participant, with each round presenting one of six types of structures. These structures vary not only in their configuration but also in the location of interventions. The specific conditions of the task are detailed in Section 3 of our study. We ensure that all the DGPs presented can be fully identified using the provided datasets. The methods and types of identification are comprehensively described in section 2.1. The distinction between our methodology and those existing in the psychological literature is comprehensively reviewed in Section 2.2.

Our main findings reveal that while participants demonstrated a 52.9% average success rate in discerning causal relationships, surpassing random chance by 19%, this performance did not uniformly extend across all causal mechanisms. In scenarios requiring the synthesis of contrasting observational and interventional evidence, success rates could decrease by up to 14% for specific mechanisms. Also, participants fell short of the 100% success rate expected of an ideal rational agent with unlimited capacities. The discovered difference between treatment and control groups indicates that the intervening nature of causality is evident for our participants. Overall, the ability to accurately recognize causal relationships was more influenced by the presence of contradictory evidence than by traditional complexity factors such as the number of mutual correlations, the number of connections between variables, or the algorithmic complexity within the Data Generating Process (DGP).

2. Conceptual Background and Questions

2.1. Discovery in mixed data: problem setting

In our study, we employ the conceptualization of causal Bayesian networks (Pearl 2009; Spirtes et al. 2000) and do-calculus to describe the distributions behind the causal mechanism in both the case of observation and intervention (manipulation). We consider a problem in which two data sets (distributions) are available to the observer – observational and post-manipulated (interventional). For correct identification, we need to elucidate two key principles. The first is how variables are interconnected, explained through the concept of d-separation. And the second is what happens if we manually change the values of one of them, clarified through discrete intervention in do-calculus. To achieve this, we first introduce our notation. The *Data Generation Process* (DGP) that is responsible for these distributions can be represented as causal Bayes net. Formally, it constitutes a directed acyclic graph (DAG) $G = (V, E)$ over a set of variables (vertex) $V = \{X_1, \ldots, X_n\}$ with a set of directed edges E and a probability distribution $P(V)$ over the graph. Following Eberhardt (2007), we make the standard assumptions in this literature: Markov (all nodes are independent of their non-descendants when conditioned on their parents); Faithfulness all conditional independences in true underlying distribution P are represented in G); Acyclicity (no cycles in the graph). A *path* from vertex V_1 to vertex V_2 in a graph is a series of connected vertices, beginning in V_1 and finishing at V_2. In this sequence, each consecutive pair of vertices is linked by an edge in the graph G.

Definition 1. d-separation. *In the context of a graph G, where X and Y represent distinct vertices $X \neq Y$, and W is a set of vertices in G excluding both X and Y, the condition for X and Y to be d-separated given Win G holds true if and only if there is no undirected path U between X and Y, such that:*

1. *every v-structure like $U_n \rightarrow U_k \leftarrow U_m$ in U has a descendent in W*
2. *no other vertex in U is in W*

If X and Y are distinct $(X \neq Y)$, and both X and Y are not included in set W, then X and Y are considered *d-connected* given the set W if and only if they are not d-separated given W.

Pearl's 2009 interpretation of interventions within the DGP provides a substantive understanding of causal relationships among variables. According to him, an intervention is considered atomic when it involves the intricate process of ”lifting” the intervened variable X from its pre-existing functional mechanism $x_i = f(\mathrm{pa}_i, u_i)$. Here, pa_i denotes the graphical parents of X, and u_i represents the unobserved influences on X. The essence of this atomic intervention lies in putting X under the influence of a new mechanism that sets the value of x_i, while all other mechanisms remain undisturbed.

Translating this into causal Bayesian networks Pearl employs a representation of intervention on a variable $X \in V$ by intervention variable I, seamlessly incorporated into the causal structure G with a direct link $I \rightarrow X$ directly impacting the intervened variable. Then Pearl allows the intervention variable to adopt values from a set encompassing $\{\mathrm{idle}, \mathrm{do}(x_i)\}$, where $x_i \in X$ and $do(\cdot)$ means that certain value is assigned by manipulation.

This stands in contrast to Fisher's randomization techniques, which set a distribution over the intervened variable. In Pearl's paradigm of atomic intervention, the intervened variable is not subjected to probabilistic variations; instead, it is ascertained at a particular value. However, akin to randomization, this intervention disrupts the causal influence of normal causes on the intervened variable. This disruption manifests in the resultant ”manipulated” distribution of X, conditioned on its graphical parents (normal causes).

$$P(x_i = x_i | pa_i, I_i) = \begin{cases} P(X_i = x_i | pa_i) & \text{if } I_i = \text{idle} \\ 0 & \text{if } I_i = \text{do}(x_i') \text{ and } x_i \neq x_i' \\ 1 & \text{if } I_i = \text{do}(x_i') \text{ and } x_i = x_i' \end{cases}$$

After defining the intervention I_i for variable X_i, we can describe the manipulated distribution $P_{\mathrm{do}(x_i')} \equiv P_{man}$ using formulation of (Spirtes et al. 2000).

Theorem 1. Manipulation Theorem (Spirtes et al. 2000). *Consider a directed acyclic graph $G = \{V, E\}$, where V represents vertices and E represents edges. Let I be the set of variables in Vthat are targeted for intervention. The $G = G_{unman}$ graph corresponds to the unmanipulated distribution $P_{unman}(V)$, while the manipulated graph G_{man} is derived by altering edges: for each variable $X \in I$, the edges incident on X are removed, and an intervention variable $I_{s(X)} \rightarrow X$ is introduced. A variable $X \in V$ is considered in $\mathrm{man}(I)$ if it undergoes intervention, indicating it is a direct child of an intervention variable $I_{s(X)}$. Then*

$$P_{unman(I)}(V) = \prod_{X \in V} P_{unman(I)}(X | pa(G_{unman}, X))$$
$$P_{man(I)}(V) = \prod_{X \in man(I)} P_{man(I)}(X | I_{s(x)} = 1) \times$$
$$\prod_{X \in V \setminus man(I)} P_{unman(I)}(X | pa(G_{unman}, X))$$

Where $pa(X)$ the parents of a node X in G are such nodes denoted as Y that are $Y \rightarrow X$.

Lemma 1. First manipulated graph lemma (Eberhardt 2007). *If G is a causal graph involving a set of variables V, and G_{man} is the manipulated graph resulting from an intervention on a subset of variables $I \subset V$, then for all pairs of variables$X$, $Y \setminus \in I$, X and Y are d-separated by some set $C \subseteq V \setminus \{X, Y\}$ in G if and only if X and Y are d-separated by some $C_{man} \subseteq V \setminus \{X, Y\}$ in G_{man}.*

Lemma 2. Second manipulated graph lemma (Eberhardt 2007). *If G is a causal graph involving a set of variables V, and G_{man} is the manipulated graph resulting from an intervention on a subset of variables$I \subset V$, then for all pairs of variables X, Y where $X \in I$ and $Y \setminus \in I$, X and Y are d-separated by some set t $C \subseteq V \setminus \{X, Y\}$ in G_{man} if and only if X and Y are non-adjacent or there is a directed edge $Y \rightarrow X$ in G.*

Definition 2. Tests: orientation and adjacency. *If there are no unobserved confounders of any of the variables V in the graph (sufficiency), then there is an X-orientation test for variables $\{X, Y\}$ in case when $X \in I$, $Y \in \setminus I$, and an adjacency test for $\{X, Y\}$ if neither is $X, Y \in \setminus I$ which return the independence relations.*

Lemma 3. Mixed identification lemma. *To determine the relationships between variables X and Y, it is sufficient to perform an adjacency test and an orientation test for each pair of variables X and Y sequentially. In such a case, we call this relationship identified by a mixed test condition.* [*]

Proof. Let us perform adjacency test first. There are two possible cases then: First, if a structural adjacency test establishes that X and Y are d-separated for a certain conditioning set, then, as per 1-st manipulated graph lemma, it concludes that they are non-adjacent (no edge between), and the analysis is complete. Otherwise the second test is needed. In this context, two possible consecutive scenarios emerge: either the X-orientation test establishes that X and Y are d-separated, or it does not. If yes, then according to 2-d manipulated graph lemma, X and Y are either non-adjacent or $X \leftarrow Y$, however previous test excludes non-adjucency consequenly $X \leftarrow Y$. If X-orientation test establishes that X and Y are d-connected , then by 2-d manipulated graph lemma $X \rightarrow Y$. This encompasses the three potential structures involving two variables. The same procedure for Y. □

If we can perform only adjacency test for given pair of variable X, Y showing that they adjacent we denote this relations as $X - Y$. We can still however identify the $X - Y$ relationships if we already identified other relationships in the graph using modified SGS algorithm.

Definition 3. modified SGS Algorithm (Spirtes et al. 2000). *Assuming acyclicity and sufficiency:*

1. *Form the complete (edge between every pair of vertices.) undirected graph H on the vertex set V.*
2. *For each pair of vertices A and B, if there exists a subset S of $V \setminus \{A, B\}$ such that A and B are d-separated given S, remove the edge between A and B from H.*
3. *Let K be the undirected graph resulting from previous step. For each triple of vertices A, B, and C such that the pair A and B and the pair B and C are each adjacent in K (written as $A - B - C$) but the pair A and C are not adjacent in K, orient $A - B - C$ as $A \rightarrow B \leftarrow C$ if and only if there is no subset S of $\{B\} \cup V \setminus \{A, C\}$ that d-separates A and C.*
4. *Let K' be the partially undirected graph resulting from previous step. Orient $A - B$ edges if they are identified by a mixed test condition.*
5. *Repeat until no more edges can be oriented: a) If $A \rightarrow B$, B and C are adjacent, A and C are not adjacent, and there is no arrowhead at B, then orient$B - C$ as $B \rightarrow C$. b) If there is a directed path from A to B, and an edge between A and B, then orient $A - B$ as $A \rightarrow B$.*

Where a directed path between two vertices V_1 and V_2 is a sequence of vertices starting with $V1$ and ending with V_2 such that for each pair of consequent vertices X_1, X_2, occurring in that order in the sequence, there is an edge $E_{X_1,X_2} = X_1 \rightarrow X_2$ in graph G.

Definition 4. Tests by contradiction. *If for any pair of variables X, Y after performing modified SGS Algorithm their relationship can be determined we call this relationship identified by a directing edges condition, and applying modified SGS Algorithm – test by contradiction.*

2.2. Related literature

Causality, particularly in its deterministic guise, represents a sophisticated cognitive process that many animals do not exhibit. This complexity is further amplified when causality intertwines with probabilistic reasoning, a combination often observed in social dynamics. Since the late 1990s, psychologists have been intrigued by how effectively humans grasp this concept in everyday scenarios. A significant body of research, notably (Bramley et al. 2017; Griffiths and Tenenbaum 2009) series of works, focuses on how individuals perform statistical inferences based on data from contingency tables. In these studies, participants are tasked with assessing the association strength between two (occasionally three) binary variables, with variations in sample size and association strength.

However, these studies only peripherally address causality. The primary reason is that the tasks do not involve learning intricate structures; instead, causality is implied through contextual narratives (for instance, participants are informed that they are analyzing data from a mouse drug trial). Consequently, the explanatory models developed in these studies are predicated on evaluating the evidence strength and its alignment with Bayesian inference. They do not, however, provide insights into which structures are more or less challenging to learn. This gap in understanding underscores the need for

[*] A nuanced detail arises here: if the intervention is atomic, as articulated by Pearl and as adopted in our design, the orientation test might not consistently yield conditional independence relations. This is because the manipulated distribution lacks certain information when the intervention assigns only one value out of several possible values. Nevertheless, it does return pairwise independence relations for the selected value under manipulation. This information can be effectively combined with an adjacency test to discern between direct and non-direct causality, essentially equivalent to establishing a conditional independence relation. Consequently, these conditions are addressed sequentially in the proof starting from adjacency.

further exploration into how humans comprehend and apply causal reasoning in various contexts particularly in causal discovery.

Another area of research integrates causal inference with active learning (Gong et al. 2023). This field primarily investigates not the interpretation of accumulated evidence in support of one model of the world over another, but rather how inferences are made in real-time as events unfold dynamically. Essentially, it seeks to understand how conclusions are drawn while observing or interacting with a process in progress and it is largely aimed at researching how people learn in everyday situations including learning among children (Schulz and Gopnik 2004). In this context, the concept of learning causal structures is already well-established. This includes scenarios where participants can intervene in a variable during a trial (Coenen et al. 2015) or across a block of trials (Steyvers et al. 2003).

Our framework, however, presents three notable distinctions from this line of research. Firstly, in these studies, the emphasis is predominantly on learning, as data is presented sequentially and participants have the freedom to choose their points of intervention. This approach makes it challenging to discern whether errors arise from imperfect learning or flawed causal reasoning. In contrast, our framework focuses exclusively on inference, deliberately excluding a learning component. This shift allows for a clearer analysis of causal reasoning capabilities. Secondly, in the existing literature, data is generated randomly, leading to varying strengths of evidence and initial signals about the Data Generating Process (DGP) for each participant. Our approach differs since it involves tasks where subjects interact directly with a statistical population, rather than with samples. This method provides a more uniform basis for evaluating causal inference. Lastly, in the active learning literature, where observations (or trials) cannot be reorganized, it is nearly impossible to identify conditionally independent relationships. These relationships are crucial as they enable the observer or actor to engage with the entire structure, rather than just analyzing pairs of variables in isolation and they are explicitly incorporated in our framework.

Taylor and Ahn's 2012 study, the most closely related to ours, also uses a static framework displaying both grouped and ungrouped data (contingency tables and direct observations). This setup enables participants to understand the origins of the contingency table and to accurately estimate the conditional probabilities necessary for proper identification However, there are notable differences between their approach and ours. Taylor's study limits choices to eight predetermined Data Generating Processes (DGPs), without the option to select individual edges, a feature our framework offers. Additionally, their research focuses on a single true DGP, whereas our framework includes nine different types, allowing for a broader exploration of causal inference. Unlike our study, Taylor's does not incorporate interventions, relying solely on the superposition of choice options with conditional independence test results for causal identification. This approach is more restrictive compared to the comprehensive methodology we employ in our research.

The focus on this topic within the economics literature is relatively limited. Spiegler (2020) provides a comprehensive review of a series of theoretical papers that delve into decision-making scenarios where agents have already selected an incorrect Data Generating Process (DGP). Our paper complements this by potentially offering insights into the nature of these decision-making errors. In terms of experimental research, our awareness is limited to a single study by economists on causality, as cited in (Kendall and Charles 2022). However, our research diverges significantly from the approach taken in (Kendall and Charles 2022). While their study examines the impact of persuasive false narratives on subjects' perceptions, it employs a DGP with only a single connection between variables and does not incorporate interventions. Our research stands out by moving away from this linear narrative approach, adopting a more structured pattern of inquiry. Similar to (Kendall and Charles 2022), our work aligns with the principles of cognitive economics (Caplin 2023), focusing on decision-making processes rather than modelling the entire spectrum of causal cognition. This distinction highlights the unique contribution of our research on exploring the complexities of causal inference and decision-making within an economic context.

In summary, our study treats the identification of cause-and-effect relationships between variables (causal discovery) as an independent task. This task precedes the process of inference and either runs concurrently with learning in continuous settings akin to everyday life situations, or operates independently of it. Our approach merges the concepts of complexity and procedural decision-making, as discussed in (Oprea 2020), with the principles of causal discovery outlined in (Eberhardt 2017). This integration is aimed at creating a well-structured task where all Data Generating Processes (DGPs) can be fully resolved. This framework allows for a comprehensive understanding of causal discovery as a distinct and crucial phase in the broader context of decision-making and causal reasoning.

2.3. Complexity and Hypothesis

In formulating the hypotheses and delineating the main design elements, let us consider the reasons behind the exploitation of a mixed dataset and how allocation to treatment and control groups aligns with our research question. The research question could be simplified to a simple distinction between observational and experimental data types as the leading factor for rigorous inference, were it not for the problems associated with its asymmetric nature. Consider the fact that inferences drawn from purely observational data can suggest the presence of an effect but fall short in determining its direction. This limitation implies that purely "observational conclusions" are inherently incomplete, lacking the necessary information to identify the direction of the effect (Eberhardt 2017). Consequently, already in the case $X \rightarrow Y$ of two variables, an asymmetry emerges: if an effect is observed after an intervention on X, it is also seen in observational data. Therefore, one can form an understanding of how the intervention influences X on Y without additional information. Conversely, if the relationship is reversed, the presence of the connection remains traceable excluding orientation of links in the observational dataset, but after intervening on X, Y remains unchanged. Consequently in this scenario, to accurately determine the relationship and its direction, it is necessary to consider two conditions at once. When dealing with more than three variables connected with incoming links from an intervention variable, interventions do not invoke changes in only one variable, leading to the issue of direct and mediated causation. Yet here the links are unidirectional, and the number of options is limited to either $X \rightarrow Y \rightarrow Z$ or $X \rightarrow Z \rightarrow Y$. In a multivariable scenario with connected variables, there is only one case where observational data could lead to precise identification without using information from interventional data—when there is a collider ($X \rightarrow Y \leftarrow Z$). However, even though this is possible, it still requires making all pairwise comparisons, imposing a considerable load on working memory in the same manner (two conditions or more).

The foundational idea that observed data and intervention data demand inference with different computational efforts forms the basis of our design. Our approach involves comparing tasks with mixed datasets representing the same causal mechanism, where the observational data remain constant, but the interventional dataset differs. In particular, it differentiates based on the placement of the interventional variable. Given that inference from pure observational data is challenging and the data generation process can be reconstructed through observations only up to its observational equivalent, we use at least one representative Data Generating Process (DGP) from each class of observational equivalence for 3-variable structures. There are only three DGPs where correct comparisons are possible in 3-variable structures, provided that the structure and number of interventions are consistent and simultaneously sufficient for accurate identification: $X \rightarrow Y \leftarrow Z$; $X \quad Y \leftarrow Z$ and $X \rightarrow Y \rightarrow Z$ (see fig. 2). Hence, we can assess whether the Data Generating Process (DGP) identifiable predominantly by intervention data yields superior recognition by comparing the outcomes of identifying these structures in the treatment and control groups (see hyp. 2 in subsec. 2.3.1 for details). Additionally, considering that all tasks of this type appear uniform (participants observe data from two tables with three columns), we enhance our data gathering by having participants identify several Data Generating Processes (DGPs) consecutively. Success rates in this approach would reveal true recognition abilities only if successes in subsequent tasks do not influence each other. This design allows us to present participants with a set of additional DGPs (with common data sets for treatment and control groups) to assess the additional influence of DGP structure on performance, using within-subjects assessment (with lower results reliability compared to assessing the direct difference between treatment and control groups, see subsec. 2.3.2 for details). The distinction between treatment and control groups in this setting would not only demonstrate that individuals grasp the crucial role of interventions in establishing causal relationships but also indicate that this task is intricately tied to the computational ability of the agent, aligning with the perspective championed by classic Herbert Simon.

2.3.1. General hypothesises (between the subjects). First of all, we hypothesize that the overall accuracy in recognizing all DGPs under consideration would be relatively low in comparisons of a rational guessers, indicating the complexity of the task at hand.

Hypothesis 1. bounded-rationality. *The percentage of correctly identified DGP by a bounded rational agent would be significantly lower than the 100% accuracy achieved by a fully rational agent, in all cases identified either by the mixed test condition (see Lemma 3) or by the directing edges condition (refer to Def. 3).*

In our second assumption, we suggest that it will be more straightforward for subjects to ascertain the existence and direction of a connection using interventional data, as opposed to observational data. This complexity arises because test results based on observational data may conflict with results of test from interventional data, necessitating a more nuanced analysis. This analysis involves considering two key conditions: the existence of an adjacency (from observation) and the impossibility of a reverse connection (from the absence of intervention results). Thus for all DGPs included in the control condition (see fig. 2) the orientation test is enough while the treatment condition additionally demands either adjacency or orientation test.

Hypothesis 2. treatment-control condition. *The ability of individuals to identify connecting edges between two variables, X and Y, in a graph G depends on the type of evidence. Independently of whether the effect is traceable from observational data or after the intervention, if identifiability and the remaining structure persist, then this dependency is based on whether the manipulated graph includes $X \in I$ and $Y \rightarrow X$ or $Y \in I$ and $Y \rightarrow X$.*

2.3.2. Complexity hypothesises (within the subjects). Following Oprea, we will consider our benchmark complexity simply as the number of relationships between variables.

Hypothesis 3. t-complexity (Oprea 2020). *The DGP represented by graph G_c is more complex than DGP represented by graph G_s (simple) if G_c has more edges (originally transitions) in excess of states than G_s* $(|V_c| - |E_c| > |V_s| - |E_s|)$

Taking into account the misleading nature of correlations, we hypothesize next that the total number of correlations associated with each node in the graph will determine the level of difficulty faced by the participant.

Hypothesis 4. c-complexity (correlations). *The DGP represented by graph G_c is more complex than DGP represented by graph G_s (simple) if G_c has more total correlations (pairwise independence test results without conditioning) among nodes than G_s.*

Hypothesis 5 is intricately linked to the principles of algorithmic complexity and the concept of Markov equivalence classes in directed acyclic graphs (DAGs). A Markov equivalence class encompasses a group of DAGs that represent identical sets of conditionally independent relationships. Consequently, when relying solely on observational data, it becomes impossible to differentiate between DAGs within the same equivalence class due to their observational indistinguishability. The complexity is determined by the length of the shortest possible description that can accurately represent these connections, which corresponds with the classical algorithmic definition described as the amount of information necessary to succinctly describe a string or a dataset. The greater the number of conditions required and the fewer the DGP states that are distinguishable under these condition combinations, the higher the resultant complexity. Formally:

Hypothesis 5. i-complexity (identification). *The ease of recognizing an edge in a graph is proportional to the fewer number of tests (adjacency test, orientation test, and directing edges condition) required for its identification.*

The following hypothesis in line with our division into a treatment and a control group suggests that the difficulty of identifying an edge depends on the type of dataset and type of test result from that dataset required for consideration. We consider a test result positive if it shows changes in the distribution of one variable given the values of another and negative if it demonstrates the absence of such changes. Then:

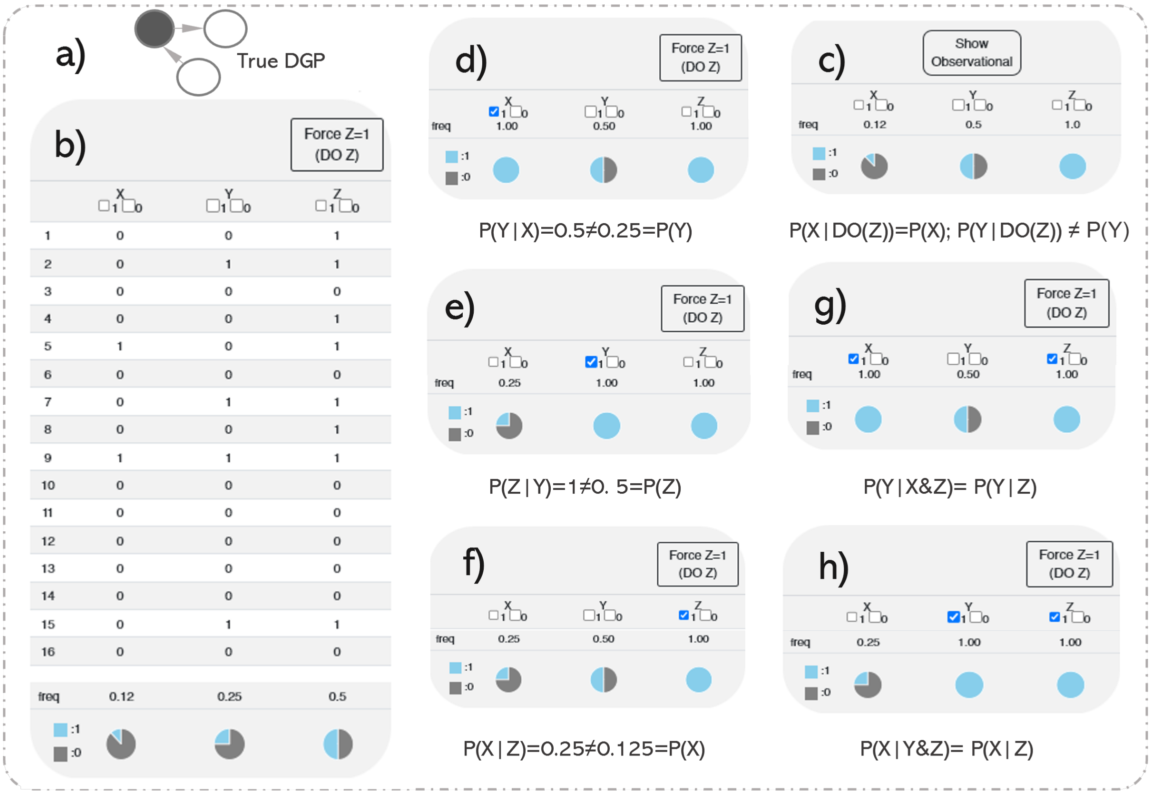
Hypothesis 6. s-complexity (superposition). *The sequence from lower to higher complexity is as follows: Initially, those edges go, which are identified either through the positive orientation test or the pairwise negative adjacency test. Subsequently, are coming edges identifiable by a combination of the adjacency test, factoring in the conditional set. Next, edges – identified concurrently by the positive orientation test and the negative adjacency test. Finally, the most complex identification involves using a positive adjacency test in conjunction with a negative orientation test.*

Based on this hypothesizes, the accuracy of identifying edges with simpler complexity should not exceed the accuracy of identifying edges with greater complexity for each pair of n complexity levels.

3. Design

3.1. Experimental task: DGPs and datasets

In our approach, the decision maker analyzes two datasets: an observed dataset (as shown in Figs 1,7 b) and an intervention dataset (depicted in Figs 1,7 c). Each dataset contains 16 observations/trials for three binary variables and allows to show of filtered (conditioned) data. These datasets represent a general population, not a sample, so the frequencies directly reflect the dependent and independent relationships between variables. We induce this understanding in participants, as explained in the procedure section.



	X □1 □0	Y □1 □0	Z □1 □0
1	0	0	1
2	0	1	1
3	0	0	0
4	0	0	1
5	1	0	1
6	0	0	0
7	0	1	1
8	0	0	1
9	1	1	1
10	0	0	0
11	0	0	0
12	0	0	0
13	0	0	0
14	0	0	0
15	0	1	1
16	0	0	0
freq	0.12	0.25	0.5

P(Y | X)=0.5≠0.25=P(Y)

P(X | DO(Z))=P(X); P(Y | DO(Z)) ≠ P(Y)

P(Z | Y)=1≠0. 5=P(Z)

P(Y | X&Z)= P(Y | Z)

P(X | Z)=0.25≠0.125=P(X)

P(X | Y&Z)= P(X | Z)

Figure 1. The data-sets representations a) true DGP b) full observational data-set without conditioning c) full interventional data-set without conditioning d)–h) conditioning under observational data set. Here, for the purpose of illustration the tables b)–h) show only frequencies for full view see fig. 7. In panel b, the interface is presented as seen by the participant. The interface includes a table displaying data and two available actions. With one action, the participant can observe changes in the frequencies of other variables. By filtering (conditioning) one variable to a value of 0 or 1, the participant can view the results presented in panels d-h. The other available action allows switching between datasets (observational and interventional) with the intervention on the variable above the pressed button, revealing what is shown in panel c. The participants can return back to the previous dataset by clicking the "show observational" button.

To solve the problem, the decision maker needs to be allowed to perform several types of tests, this possibility is exemplified in Fig 1, where a mediation DGP (a chain) with mediator variable Z is presented. The Adjacency Paired Test, for example, is demonstrated in the connection between X and Y. The test is shown by comparisons in Fig 1, panels b) with e) or d), where a frequency change establishes that X and Y are connected directly, through Z, or both (e.g., in b) $d) \dashv P(Y|X) = 0.5 \neq 0.25 = P(Y) \vDash e)$).

The Adjacency Given Conditional Set Test (Conditional Independence Test) for variables $X$and Y given Z is exemplified through comparisons of f) and h), where$P(X|Y\&Z) = P(X|Z) = 0.25$, indicating no direct connection.

The orientation Test is illustrated by comparing panels b) and c). Following the concept of intervention as do(var) from Spirtes et al. (2000), here $P(X|DO(Z)) = P(X); P(Y|DO(Z)) \neq P(Y)$, indicating a connection $Z \rightarrow Y$, and no directed edge from Z to X.

By applying these tests to each pair of variables, we reveal the structure $X - Z \rightarrow Y$. To identify the last $Z - X$ direction we can either use a Test by Contradiction or state that the $Z \leftarrow X$ connection is identified by a Mixed Test Condition.

Through this methodology, a rational decision-maker can completely identify the original DGP. In our study, we employ six different types of Data Generating Processes (DGPs), each involving the three variables depicted in Figure 2. Each DGP is characterized by a unique set of conditional independencies, with the exception of the ”fork” and ”two links” configurations, which share an identical set. In these scenarios, the intervention dataset will differ between the treatment and control groups for the three DGPs.

We specifically modify only these three DGPs because identification in others is feasible only if we alter the frequencies or the number of interventions. It’s important to note that all these DGPs can be successfully identified using the observational and interventional datasets, provided the tests described earlier are applied.

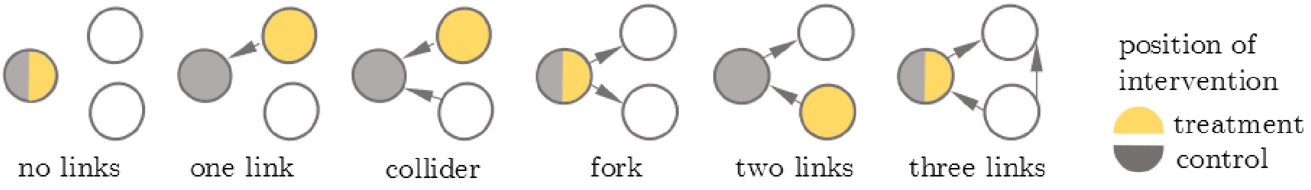


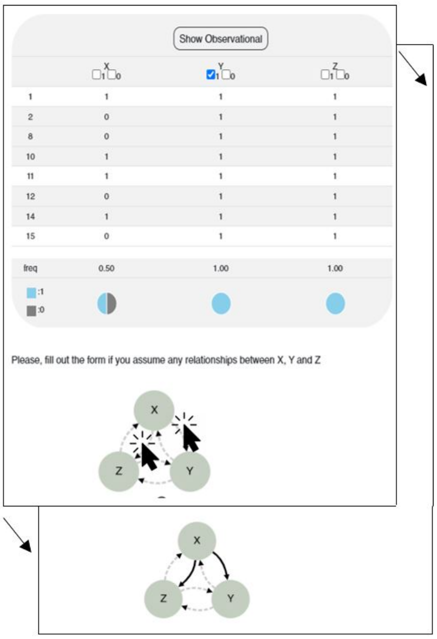
Figure 2. Types of data generation process in treatment and control groups

3.2. Experimental Procedure

In our framework, we assess the ability to identify causal relationships directly from factual information. This is aimed at understanding decision-making under flawed causal reasoning. Participants are asked to solve problems of varying difficulty, based on the complexity of different Data Generating Processes (DGPs). They repeatedly observe data and identify the underlying DGP, as shown in Fig. 3. Each round presents data sets in three columns with 16 rows each, representing binary values determined by the system’s random choice, but frequency depends on the mechanism assigned to each column. The interface allows two interactions: generating an interventional sample for one variable and filtering each variable by value. The dependency of column pairs through their mechanisms is unknown to participants.

Participants are informed about the decision-making process and potential rewards, emphasizing the importance of careful instruction reading. The experiment’s anonymity and confidentiality are assured, with tokens used as experimental monetary units.

The task narrative (see supplementary for the instructions details) involves the Little Prince exploring plant growth on different planets, with seeds of three types (X, Y, Z) in various maturity states. The task is to determine if one plant type aids another’s growth when planted side by side [†], considering pseudo-randomness[‡] and initial distributions [§]. Each planet has a twin with identical properties. For inducing the understanding of the interventional nature of data it was told that: *“To determine the kind of dependencies between seeds on planets of each type, the Little Prince*



	X ☐1 ☐0	Y ☑1 ☐0	Z ☐1 ☐0
1	1	1	1
2	0	1	1
8	0	1	1
10	1	1	1
11	1	1	1
12	0	1	1
14	1	1	1
15	0	1	1
freq	0.50	1.00	1.00
🟦:1 ⬛:0	◐	●	●

Please, fill out the form if you assume any relationships between X, Y and Z

Figure 3. Example of interface within one round

[†] We use neutral language and determine cause and effect relationships as ”weak-strong relationships” within the task. We are also simplifying tracking by reducing influence to the monotonic: the positive event in one variable could induce only the positive event in another.
[‡] Different soil types: sandy or rocky (not)allows weak-strong relationships to work
[§] We introduce it through the presence of mature seeds that always growth (becoming equal to 1) but visibly do not distinguishing from others

uses the following method. On each planet, he sows 16 beds with a set of the types X, Y, and Z. On the corresponding twin planet, within each bed, he replaces a seed of a certain type with a mature seed of that same type from the container. He plants these mature seeds after the others so that all seeds that would naturally grow without the impact of the chosen seeds have already done so. That way other strong seeds will no longer be able to influence it, even if it is weak towards them. However, it will be able to affect those seeds that are weak in relation to it, and those in the chain afterward." The goal in this task is to deduce the weak-strong (we do not use "cause-effect" to avoid framing) dependencies between seeds on each planet type based on growth patterns in 16 beds, comparing each planet with its twin.

The session structure involves solving 18 typical tasks (each of 6 types of DGP is shown once in one of each block 1-6, 7-12, 13-18 round), with correct answers starting to be shown after the 6th task. Tokens are awarded based on accuracy in identifying links, with a maximum penalty of 9.75 tokens and a reward of 10 tokens per task for perfect accuracy. Penalties vary for missing links, incorrect links, and mix-ups in identifying the direction of cause-effect relationships between plants X, Y, and Z. We also asked them to report their confidence in an incentivized manner. When their actual accuracy matched with the confidence, they received a bonus, with a maximum value of 5 tokens per round. Thus maximum payoff in the round was equal to 15 tokens.

3.3. Experiments' description

Experiments were conducted in 2023 at the Laboratory for Experimental and Behavioral Economics at the Higher School of Economics (HSE), Moscow. It used oTree (Chen et al. 2016), an open-source platform for running laboratory, online, and field experiments. Participants in the experiment were compensated using tokens, which served as experimental monetary units. The amount of tokens awarded to each participant depended on their accuracy in solving the tasks. For each task, if all connections were correctly identified, the participant received 15 tokens per round (45 ₽). We collected observations of 2592 decisions from 48 participants in totall, see Table 1 for sample description.

Table 1. Sample description and payoffs

Condition	N	Share of male	min age	max age	mean age	share of correctly found edges	mean payoff
controll	27	0.51	18	43	26	0.526	613 ₽
treatment	21	0.52	18	46	25.3	0.534	617 ₽

4. Results

4.1. General observations: seeing versus doing

The experimental findings are detailed in Table 2, offering descriptive statistics, and are graphically depicted in Figure 4. Table 2 elucidates the discrepancy between mean and median accuracy as well as between participants' perceived confidence and their actual accuracy listed by DGPs. This observation reveals that confidence levels are consistently higher with one exclusion than actual accuracy, which is in line with the established patterns observed in the overconfidence literature.

Table 2. The descriptive statistics groped by data generation processes

DGP	N	mean confidence	mean accuracy	median confidence	median accuracy	confidence st. deviation	accuracy st. deviation	minimal accuracy	max accuracy
collider	48.00	59.17	0.50	58.33	0.44	20.81	0.20	0.00	1
fork	48.00	57.60	0.58	60.00	0.56	22.77	0.21	0.00	1
nolinks	48.00	57.85	0.61	59.17	0.56	23.45	0.34	0.00	1
onelink	48.00	58.68	0.58	58.33	0.61	22.68	0.28	0.00	1
threelinks	48.00	61.63	0.40	65.00	0.44	21.05	0.19	0.00	0.77
twolinks	48.00	60.52	0.51	62.50	0.56	21.53	0.18	0.22	1

Figure 4 provides a comprehensive summary of the results, organizing them by treatment/control conditions, Data Generating Processes (DGPs), and experimental blocks (with each block featuring at least one instance of a DGP). In this figure, the average accuracy levels of subjects are illustrated with bars, while individual performances are denoted by dots. A notable insight from this analysis is that feedback provided in the second and third blocks, along with the recurrent appearance of DGPs, does not markedly affect accuracy levels. In the treatment condition with the "two links" DGP, there even appears to be a slight decrease in accuracy. This is also evidenced by the fact that none of the pairwise tests of means comparison retain significance after correction for multiple comparisons. It also means that we can utilize the entire set of observations in further analyses without looking back for potential learning effects.

The data, both at the subgroup level in the figure and the aggregate level in the table, indicate that participants typically perform better than a random guesser (expected accuracy of 0.33). It becomes apparent that the type of data generation process plays a role in determining accuracy levels. Notably, accuracy demonstrates variability across different DGPs, at

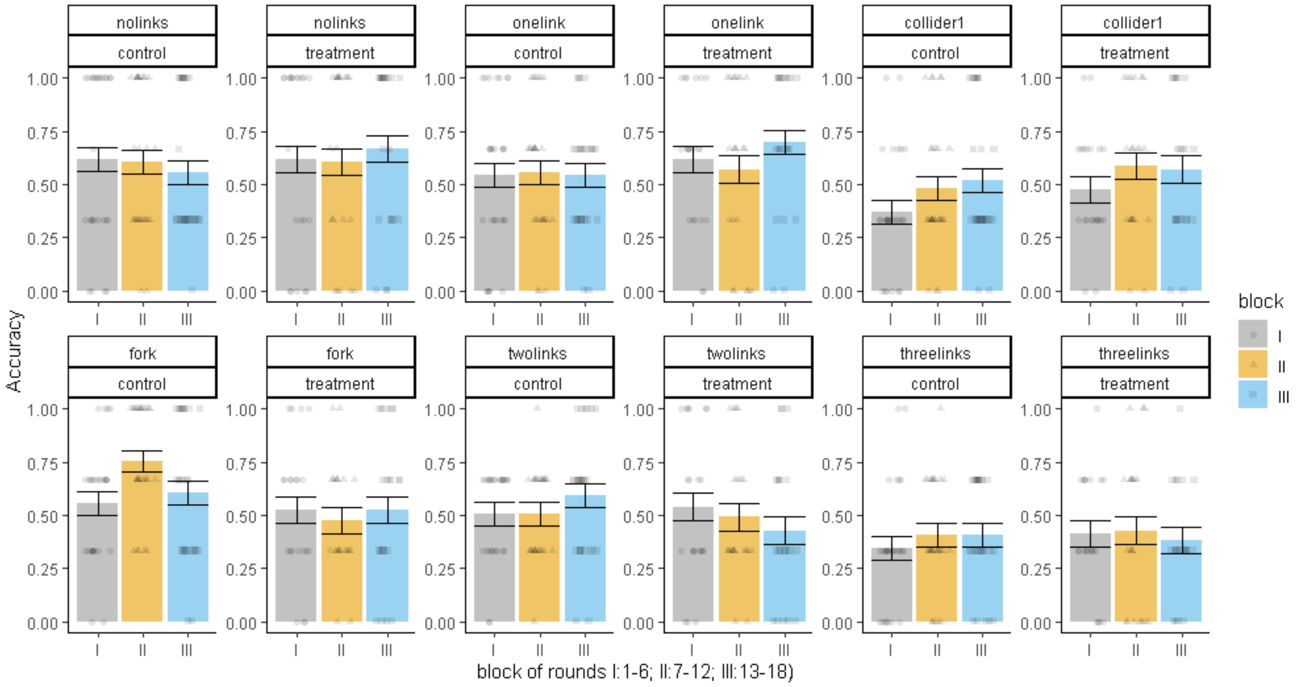


Figure 4. The general observations grouped by blocks, treatment conditions, and DGPs. The random guesser accuracy is 33%.

least partially resulting in significantly lower performance with more complex DGP. For example, in the ”three links” DGP scenario, no participants successfully identified all three connections concurrently. This leads to our initial result:

Result 1. *In our task of discerning causal inter-variable dependencies from provided evidence, participants demonstrate an average success rate of 52.9% (95% confidence interval ranging from 0.51 to 0.54 in the binomial test). This rate exceeds random chance performance by 19% (p<0.0001), however, this result falls notably short of the benchmark set by a rational agent, who would be expected to attain a 100% success rate (p-value <0.0001).*

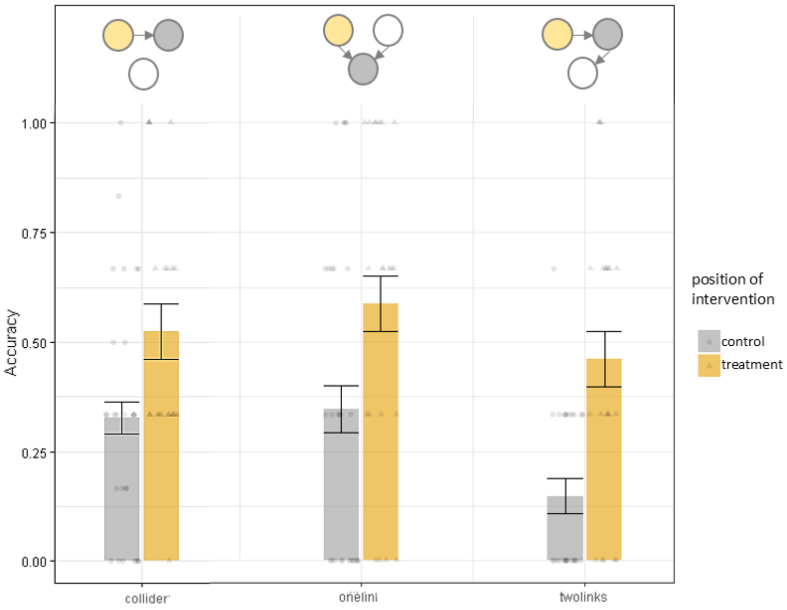


Figure 5. The difference between accuracy in revealing the ”observational” edge in the control group and ”interventions” edge in the treatment group (edge between orange and grey nodes). The random guesser accuracy is 33%.

The observation that participants’ performance is 19% better than a random response naturally raises a subsequent question: To what extent does this improvement stem from the nature of the evidence participants are engaging with? Specifically, is the enhancement in performance attributable to active observations of plant growth following the prince’s interventions, or to passive observations of plant growth under default conditions? Our experimental design was tuned to quantify this distinction, taking into account both the presence and absence of the influence exerted by the Data Generating Process (DGP) structure. The outcomes of these assessments are concisely summarized in Figure 5. The analysis reveals a notable disparity in the recognition of causal relationships based on the type of evidence available. Specifically, edges that are inferred under the condition of observable effects, but in the absence of intervention, are noticeably less accurately identified compared to those discerned from intervention effects, considering the rest of the graph structure. The impact of the graph's structure type is also evident: the difference in precision is 19.6% for a collider, 24.2% for a "one link" and as much as 31.1% for "two links". The mean effect size across these conditions is 23.6%, with this difference being statistically significant (p-value < 0.0001 in the Wilcoxon test with Bonferroni correction). Therefore, our subsequent result can be formulated as follows:

Result 2. *In cases involving three variables, the task of correctly uncovering true dependencies proves to be significantly more challenging when the provided evidence of an effect is apparent only through observational data, compared to cases where the effect is traceable exclusively from interventional data.*

4.2. Which complexity is better

In the preceding subsection, we briefly touched upon the variations in accuracy that arise from the structure of the Data Generating Process (DGP). We are continuing this consideration in detail in the present section. The notable 21% disparity in accuracy between the 'three links' and 'no links' scenarios following from the table 2 is evidently a consequence of their differing structural complexities. This leads us to a conceptual divarication in our analysis: Does the overall structure of the DGP predominantly influence accuracy, or do similarities within specific elements across different structures play a more significant role than their immediate connections within the DGP?

If the overall structure is the primary influencer, then the most appropriate metric for description would be t-complexity, which delineates the boundaries of the structure. On the other hand, if it is more effective to dissect the DGP into segments and categorize these segments based on distinct characteristics, then the selection of these characteristics should align with hypotheses 4-6. This approach would imply that understanding the DGP and its impact on accuracy requires a more granular analysis, focusing on individual components and their interrelations within different structural frameworks.

Figure 6 presents four distinct panels, each categorizing the accuracy of task performance into groups that correspond to varying levels of complexity. These complexity levels are associated with different measures of complexity. Within each panel, individual data points, represented by dots, indicate the accuracy observed for each participant. Additionally, figures within the panels represent the estimated density distribution of these observations. The complexity levels are arranged horizontally within each panel, progressing from less difficult on the left to more difficult on the right. This arrangement allows for a nuanced visualization of how task performance accuracy correlates with different measures of complexity in the task structure.

When conducting a visual analysis of Figure 6, two key considerations are pertinent. Firstly, the centrality of the measure is crucial. A more centered distribution of data points within each complexity level indicates better performance consistency across subjects. Conversely, thicker tails in the distribution suggest greater variability in performance, implying that the task might be more challenging or less consistently understood by different subjects. Secondly, the focus should be on the position of the 'center of mass' for each complexity level. In a simplified approach, this can be approximated by the average of the data points within each category. Ideally, the center of mass for each successive level should be lower than the one preceding it. This descending trend would visually represent the increasing complexity of each category, reflecting the notion that each subsequent category is more complex and, therefore, more challenging to solve successfully than the one before it. This analysis helps to distinguish which complexity conceptualization corresponds with task difficulty more accurately.

The panels analysis reveals that both c-complexity and t-complexity are less effective in describing the data, as they fall short of meeting both of the established criteria.

For c-complexity, we observe an inconsistency in the expected trend: level three is positioned lower than level four, which contradicts the anticipated progression of complexity. Additionally, some of the distributions within c-complexity do not visually appear to be unimodal, suggesting a lack of a clear central tendency and indicating variability in how subjects respond to tasks of a given complexity level. In the case of t-complexity, there is a notable issue with the data spread, particularly in the DGP with two links (especially evident in comparison with the collider scenario). Additionally, we observe a significant dispersion, indicating a wide range of performance outcomes, especially for the "one link" scenario. These observations suggest that both c-complexity and t-complexity may not be the most suitable measures for capturing the nuances of task performance in relation to the complexity of the data generation processes in this context.

The evaluation of remaining i-complexity and s-complexity as potential candidates for better describing the data presents a more nuanced challenge, as choosing between them is not straightforward.

i-complexity demonstrates a clear advantage with its well-centered distributions and a relatively compact description encompassing only three levels. This characteristic of i-complexity suggests a more consistent understanding and performance by participants across different complexity levels. However, it does not entirely capture the increasing difficulty trend that one would expect with escalating complexity.

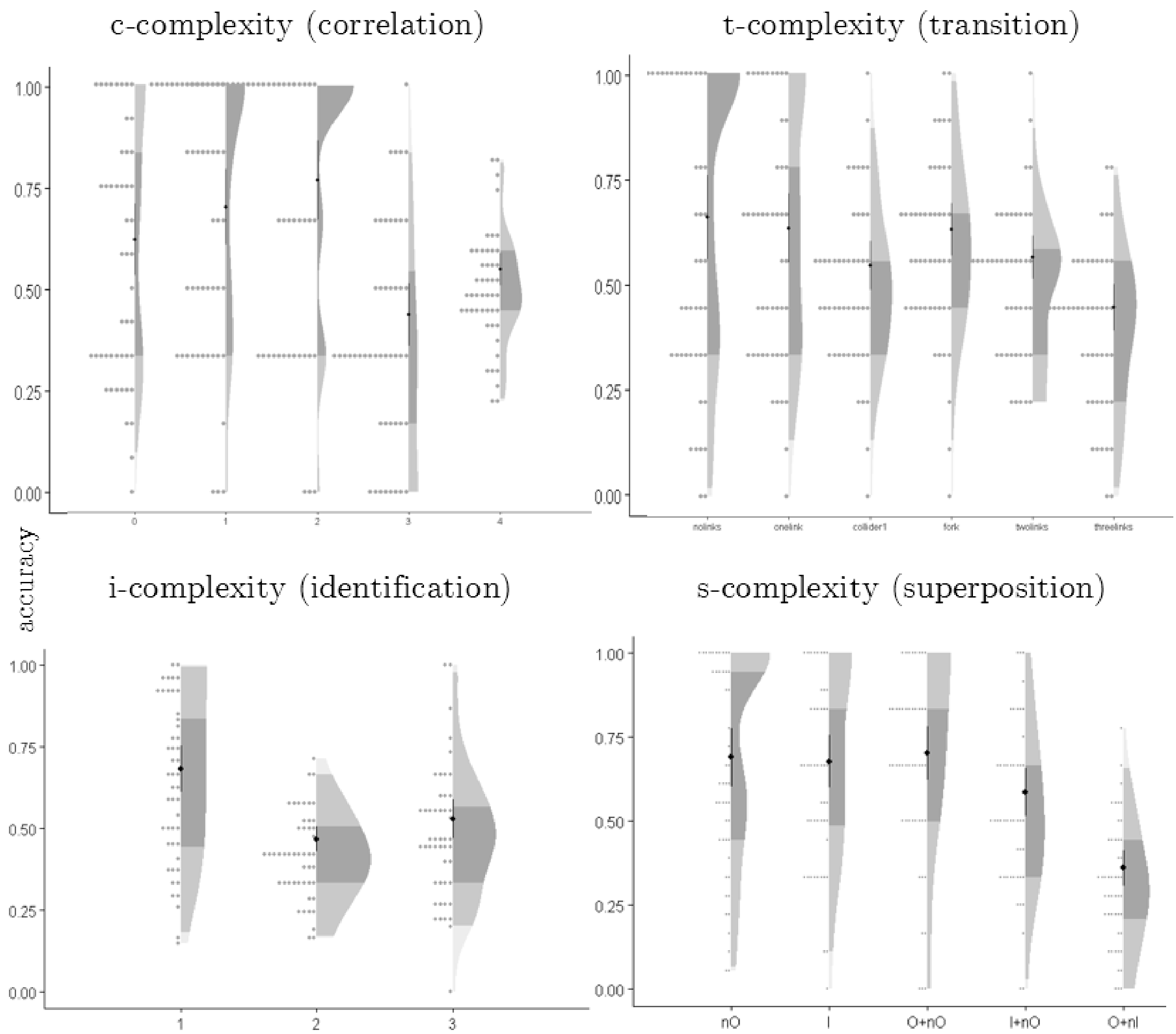


Figure 6. Distribution of accuracy of participants according to different complexity measures (hypothesizes 3-6). The difficulty levels within each panel are ordered from left to right from less difficult to more difficult. Specifically, in panel c, the numbers indicate the total amount of correlations for the nodes of a particular edge. In panel i, the focus is on the minimum number of tests required for accurate identification. Lastly, in panel s, the symbols have specific meanings: 'O' represents observed data, 'I' denotes intervention, and a lowercase 'n' signifies the absence of an effect (if it is not present).

On the other hand, s-complexity stands out as the only measure that consistently exhibits a decreasing trend in performance as complexity increases. This trend aligns with the intuitive understanding that more complex tasks should correspond to lower performance accuracy. Therefore, s-complexity seems to reflect the actual tasks complexity rather than just the similarity of categories within the data.

A potential issue with i-complexity is its inability to explain why tasks with fewer conditions appear more difficult for participants than those with more conditions unless the specific nature of these conditions is considered. This inconsistency could be a significant drawback in using i-complexity as a reliable measure.

To address the potential influence of varying abilities among participants within the sample, we employ regression analysis in addition to the conducted analyses. This approach allows for a more comprehensive understanding of how different complexity measures correlate with task performance while accounting for individual differences in participant abilities. By doing so, we aim to isolate the effect of task complexity on performance from the level of participants' abilities. The use of Multidimensional Item Response Theory (MIRT) with a two-parameter logistic (2PL) model in our study is grounded in its ability to effectively differentiate between participant ability and task difficulty (tab. 3). This approach is particularly suitable for our research, as it allows us to isolate the impact of task complexity on performance. The 2PL component of the model provides a nuanced understanding of how each task's difficulty (coefficient b) differentiates between participants of varying abilities (coefficient a). Additionally, the multidimensional nature of MIRT aligns well with our study's focus on multiple complexity dimensions, enabling the establishment of between-item dimensions as different types of complexity behind performance.

Table 3. Multidimensional two-parameter logistic item response regressions with complexity as between-item dimensions

Model	AIC	BIC	CAIC	RMSEA	EAP Reliability	Mean b in dim. 1	Mean b in dim. 2	Mean b in dim. 3	Mean b in dim. 4	Mean b in dim. 5
t-complex	3334	3547	3661	2942	0.88, 0.847, 0.83, 0.641	-1.11	-0.45	-0.18	0.56	-
s-complex	3193	3440	3572	2780	0.895, 0.715, 0.691, 0.732	-0.89	-0.82	-0.03	0.95	-
i-complex	2748	2945	3050	2364	0.87, 0.67, 0.736	-0.85	0.41	0.16	-	-
c-complex	3279	3500	3618	2884	0.878, 0.88, 0.85, 0.68, 0.82	-0.57	-0.75	-1.06	-0.22	0.0004

When we look at the regression results from Table 3, two aspects stand out that are in line with our previous analysis: the information criteria and the average b-coefficient across dimensions. The information criteria tell us how well our model explains the variations in observed performance – the lower this value, the better the model fits our data. As for the b-coefficients, they give us an idea of the difficulty level of each item in our study. A b-coefficient around 0 means the item has an average difficulty level. If the b-coefficient is less than 0, like -1 or even lower, it suggests that the item is easier than average. On the other hand, a b-coefficient greater than 0 shows that the item is more challenging than average.

In our analysis, we observe a pattern consistent with our earlier findings, where the two complexity concepts provide a better explanation of the observations compared to other models. The i-complexity model shows the lowest values in terms of information criteria, indicating a good fit, but there's an interesting twist: the b-coefficients for the second level are still higher than for the third. On the other hand, s-complexity, which has the second-best information criteria values, displays a consistent order in b-coefficients, increasing from low to high.

The results of regressions provide two key points here. First, they allow us to rule out the possibility that our results are skewed by an imbalance of abilities among our sample participants. Second, when considering the evidence at hand, we lean towards s-complexity as the more effective model, at least in the context of a 3-variable case. This preference, however, does not dismiss the potential value of combining s and i complexities. Such combination could potentially offer a more precise conceptual framework, especially when dealing with more complex Data Generating Process (DGP) structures. This hypothesis, while promising, would need further testing and validation with data encompassing a broader range of sophisticated DGP structures.

The concluding result of our analysis highlights how humans engage with causal relationships in data:

Result 3. *The human ability to accurately recognize cause-effect relationships in a 3-variable discovery task is predominantly influenced by both the nature of the data (whether it is interventional, observational, or a mix of both) and how these data types complement each other (s-complexity). Specifically, people are more adept at discerning evidence that demonstrates the presence of an effect in interventional data, or the complete absence of an effect, compared to situations involving mixed evidence. Furthermore, when different parts of the Data Generating Process (DGP) require different types of evidence, recognition and understanding of these parts are not uniform. The most challenging aspect of causal recognition is the following: identifying the effect of X on Y is most difficult when it requires a logical synthesis of two contrasting pieces of evidence. This complexity arises when one must deduce the presence of a causal connection $X \to Y$ by combining observational data (which suggests a connection between X and Y) with interventional data on Y (which shows an absence of connection following the intervention).*

5. Discussion and conclusion

In recent years, economists have shown a keen interest in understanding the cognitive aspects of economics (Caplin 2023), specifically focusing on ways of individuals navigating through evidence and information, and the aspects of processing information in decision-making environments. For instance, (Kendall and Oprea 2022) experiment suggests that people can find algorithmic representations in data, although this task is challenging. Empirical studies in economics, such as Benjamin et al. (2018), reveal the difficulties individuals face in extracting information from probabilistic outcomes due to computational limitations.

The interest in these economic modelling of cognitive processes arises from the essential role inference tasks play in economic life and decision-making. Herbert Simon's idea of bounded rationality emphasizes the need for improved conceptualizations of how humans process information used for decisions. Causal behavioral inference, among other challenges, remains one of the most difficult tasks in decision-making (Spiegler 2023), combining elements of probability, learning, and counterfactual thinking. To directly assess human ability in contractual reasoning and address causality when working with data, we establish a novel testing framework. To isolate the effects of learning and inference from the understanding of causal inference fundamentals, we implement the causal discovery setting (Eberhardt 2007) as the basis for our task.

Our results indicate that identifying causal mechanisms from the data is challenging for participants, with an average success rate of only 52.9% across different Data Generating Processes (DGPs). While this result is 19% higher than random guessing, it falls significantly short of the 100% accuracy achievable in our task with just a few clicks: a level of performance that a professional or an agent with unlimited computational capacities would attain. In scenarios requiring the synthesis of contrasting observational and intervention data, the success rate could drop to 14% for specific places in the mechanism (chain/mediation). This surprisingly illustrates the fact that providing evidence of experimentally verified effects from the mediation/proximate variable to the termination variable can take attention away from the distal/ultimate cause in the mechanism.

The observed differences between treatment and control groups suggest that the intervening nature of causal interference for our participants is evident. However, even with this understanding, the capacity to accurately identify causal relationships depends more on the presence of contradictory data evidence than on traditional complexity factors. By traditional factors, we mean the number of mutual correlations, the number of relationships between variables (Oprea's t-complexity 2020), and the algorithmic complexity of the data generation process (DGP). This underscores that the comprehension of causal relationships is significantly influenced by the epistemic role of interventions and does not strictly align with the complexity of simple computational problems. Moreover, within a single DGP, complexity can vary.

The framework, while already complex, has limitations due to simplifications made during its development, primarily stemming from the use of Directed Acyclic Graphs (DAGs) to establish causal mechanisms. Feedback loops and cyclical causal processes common in real-world scenarios are not captured or analyzed within a DAG framework. The found pattern may not fully generalize the systems where cyclic causality is present. Additionally, we do not address issues related to the possibilities of unobserved variables that can influence the relationships between the variables being studied in an uncontrolled environment. Lastly, the independent relationships behind DGPs require specific differences in probabilities, adding an extra layer of complexity to the task that may affect participants' ability to detect differences in compared distributions. However, this only increases noise in measurement because it is equally reflected in real-life scenarios as well and cannot be completely eliminated.

Despite these limitations, the present findings have broad relevance for causal cognition literature and cognitive economics (Caplin 2023). By examining how people discern and interpret causal relationships, the study provides a comprehensive analysis of cases involving mixed data. This understanding can inform the development of educational tools and strategies to improve critical thinking and analytical skills, particularly in contexts where causal reasoning is essential. It includes cases of policy discussions, particularly in areas like taxation. For instance, in the scenario involving $\begin{tikzcd} X & Y \arrow[l] & Z \arrow[l] \arrow[ll, bend right] \end{tikzcd}$; tax rate (Z), tax base (Y), and tax collected (X), the study highlights the potential risks of oversimplified interpretations of causal relationships. If policy arguments solely rely on experimental findings suggesting that increasing the tax rate will boost tax collection, they may overlook the complex interplay between these variables. Thus on account of the present findings, the development of a more nuanced protocol for discussing public policy and regulations becomes possible. Another potential implication is in decision support systems, where understanding the effects of misinterpretation can be integrated. By recognizing common pitfalls in causal inference, these systems can be designed to provide more accurate and reliable guidance, particularly in complex scenarios where causal relationships are not straightforward.

References

Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, Bollen KA, Brembs B, Brown L, Camerer C et al. (2018) Redefine statistical significance. *Nature human behaviour* 2(1): 6–10.

Bramley NR, Dayan P, Griffiths TL and Lagnado DA (2017) Formalizing neurath's ship: Approximate algorithms for online causal learning. *Psychological review* 124(3): 301.

Caplin A (2023) *Science Of Mistakes, The: Lecture Notes On Economic Data Engineering*, volume 16. World Scientific.

Chen DL, Schonger M and Wickens C (2016) otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance* 9: 88–97.

Coenen A, Rehder B and Gureckis TM (2015) Strategies to intervene on causal systems are adaptively selected. *Cognitive psychology* 79: 102–133.

Eberhardt F (2007) Causation and intervention. *Unpublished doctoral dissertation, Carnegie Mellon University* : 93.

Eberhardt F (2017) Introduction to the foundations of causal discovery. *International Journal of Data Science and Analytics* 3: 81–91.

Galles D and Pearl J (1998) An axiomatic characterization of causal counterfactuals. *Foundations of Science* 3: 151–182.

Gong T, Gerstenberg T, Mayrhofer R and Bramley NR (2023) Active causal structure learning in continuous time. *Cognitive Psychology* 140: 101542.

Griffiths TL and Tenenbaum JB (2009) Theory-based causal induction. *Psychological review* 116(4): 661.

Kendall C and Oprea R (2022) On the complexity of forming mental models. *forthcoming at Quantitative Economics* .

Kendall CW and Charles C (2022) Causal narratives. Technical report, National Bureau of Economic Research.

Oprea R (2020) What makes a rule complex? *American economic review* 110(12): 3913–3951.

Pearl J (2009) *Causality*. Cambridge university press.

Schulz LE and Gopnik A (2004) Causal learning across domains. *Developmental psychology* 40(2): 162.

Spiegler R (2020) Behavioral implications of causal misperceptions. *Annual Review of Economics* 12: 81–106.

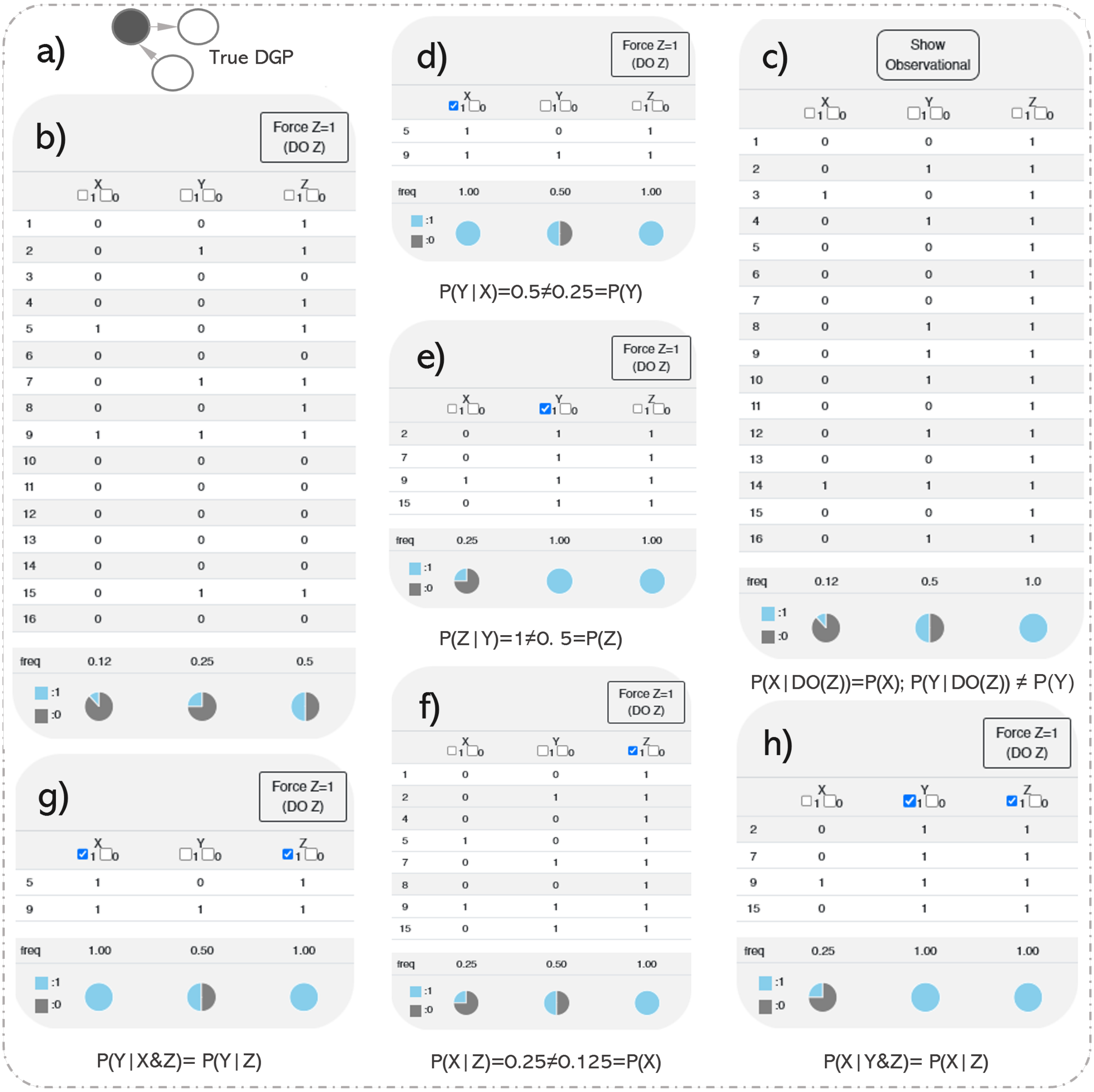
Spiegler R (2023) Behavioral causal inference. *arXiv preprint arXiv:2305.18916* .

Spirtes P, Glymour CN and Scheines R (2000) *Causation, prediction, and search*. MIT press.

Steyvers M, Tenenbaum JB, Wagenmakers EJ and Blum B (2003) Inferring causal networks from observations and interventions. *Cognitive science* 27(3): 453–489.

Taylor EG and Ahn Wk (2012) Causal imprinting in causal structure learning. *Cognitive psychology* 65(3): 381–413.

Supplementary



b)

Force Z=1 (DO Z)

	X □1 □0	Y □1 □0	Z □1 □0
1	0	0	1
2	0	1	1
3	0	0	0
4	0	0	1
5	1	0	1
6	0	0	0
7	0	1	1
8	0	0	1
9	1	1	1
10	0	0	0
11	0	0	0
12	0	0	0
13	0	0	0
14	0	0	0
15	0	1	1
16	0	0	0
freq	0.12	0.25	0.5

■:1 ■:0

g)

Force Z=1 (DO Z)

	X ☑1 □0	Y □1 □0	Z ☑1 □0
5	1	0	1
9	1	1	1
freq	1.00	0.50	1.00

■:1 ■:0

P(Y|X&Z)= P(Y|Z)

d)

Force Z=1 (DO Z)

	X ☑1 □0	Y □1 □0	Z □1 □0
5	1	0	1
9	1	1	1
freq	1.00	0.50	1.00

■:1 ■:0

P(Y|X)=0.5≠0.25=P(Y)

e)

Force Z=1 (DO Z)

	X □1 □0	Y ☑1 □0	Z □1 □0
2	0	1	1
7	0	1	1
9	1	1	1
15	0	1	1
freq	0.25	1.00	1.00

■:1 ■:0

P(Z|Y)=1≠0. 5=P(Z)

f)

Force Z=1 (DO Z)

	X □1 □0	Y □1 □0	Z ☑1 □0
1	0	0	1
2	0	1	1
4	0	0	1
5	1	0	1
7	0	1	1
8	0	0	1
9	1	1	1
15	0	1	1
freq	0.25	0.50	1.00

■:1 ■:0

P(X|Z)=0.25≠0.125=P(X)

c)

Show Observational

	X □1 □0	Y □1 □0	Z □1 □0
1	0	0	1
2	0	1	1
3	1	0	1
4	0	1	1
5	0	0	1
6	0	0	1
7	0	0	1
8	0	1	1
9	0	1	1
10	0	1	1
11	0	0	1
12	0	1	1
13	0	0	1
14	1	1	1
15	0	0	1
16	0	1	1
freq	0.12	0.5	1.0

■:1 ■:0

P(X|DO(Z))=P(X); P(Y|DO(Z)) ≠ P(Y)

h)

Force Z=1 (DO Z)

	X □1 □0	Y ☑1 □0	Z ☑1 □0
2	0	1	1
7	0	1	1
9	1	1	1
15	0	1	1
freq	0.25	1.00	1.00

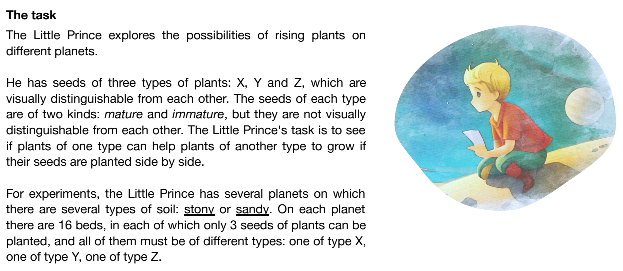
■:1 ■:0

P(X|Y&Z)= P(X|Z)

Figure 7. The data-sets representations a) true DGP b) full observational data-set without conditioning c) full interventional data-set without conditioning d)–h) conditioning under observational data-set. Decision maker within the task observes only one table with frequencies at once (with no formula). In panel b, the interface is presented as seen by the participant. The interface includes a table displaying data and two available actions. With one action, the participant can observe changes in the frequencies of other variables. By filtering (conditioning) one variable to a value of 0 or 1, the participant can view the results presented in panels d-h. The other available action allows switching between datasets (observational and interventional) with the intervention on the variable above the pressed button, revealing what is shown in panel c. The participants can return back to the previous dataset by clicking ”show observational” button.

Table 4. Difficulties (b-coef.) of each item in each regression model

Model	b_Dim1	b_Dim2	b_Dim3	b_Dim4	b_Dim5
t-complex	-1.036, -2.203, -1.816, -1.176, -0.483, -1.013, -0.639, -0.597, -1.047	-0.175, -0.506, -0.636, -0.367, -0.394, -0.348, -0.302, -0.791, -0.604	0.431, -0.168, 0.791, 0.337, 0.252, -1.172, -0.495, -0.523, -0.01, 0.083, 0.083, -0.787, -0.626, -0.742, -0.434, -0.605, -1.076, 0.28, -0.104, 0, -0.174, -0.451, -0.452, 0.887, 0.257, -0.506, -0.167	1.06, -0.351, 1.951, 0.172, 0.616, 0.268, 0.347, -0.001, 1.015	NA
s-complex	-0.811, -0.051, -0.477, -1.465, -1.088, -1.296, -0.465, -0.878, -0.138, -0.388, -0.051, -2.449, -0.779, -1.72, -0.359, -0.781, -1.071, -1.093, -0.7, -0.631, -0.848, -1.097, -0.613, -0.659, -0.711, -2.403, -1.233	-0.625, -0.723, -1.021, -0.752, -0.442, -1.37	0.247, -0.146, -0.443, -0.506, -0.294, 0.238, -0.63, 1.007, -2.065, 2.134, -0.268, 0.269	1.006, 0.773, 1.133, 1.685, 1.077, 0.953, 0.735, 1.412, 0.451, 1.093, 0.465, 0.975, 1.717, 0.89, 0.65, 0.666, 0.506, 1.003	NA
i-complex	-0.723, -1.143, -1.921, -0.863, -0.919, -0.683, -0.516, -0.264, -0.397, -0.653, -1.69, -0.999, -0.737, -0.561, -0.751, -0.719, -1.052, -0.879, -0.345, -1.016, -1.081	0.897, 0.006, -0.001, 1.287, -0.267, 0.185, 1.475, -0.167, -0.28, 1.245, 1.165, -0.708, 0.63, -0.355, 0.343, 1.144	0.078, 0.081, -0.183, -0.859, 0.082, 0.205, 0.081, 0.251, 0.299, 1.127, 0.439, -0.096, 0.091, 0.7	NA	NA
c-complex	-0.466, -1.783, -1.544, 0.331, -0.047, -1.235, -0.326, -0.493, 0.413	-0.627, -0.355, -0.984, -0.941, -1.175, -0.991, -0.185	-1.57, -1.019, -0.594	-3.237, -1.049, 1.176, 0.831, 0.373, 0.555	0.168, 0.285, 0.656, -0.094, 1.062, -0.463, 0.185, -0.103, 0.625, -0.342, -0.509, -0.303, 0.094, -0.201, -0.476, 0.007, 0.346, -0.078, -0.609, 0.547, 0.05, -0.431, 1.253, -0.083, -1.729, -0.432, 0.248, 0.084, 0.256



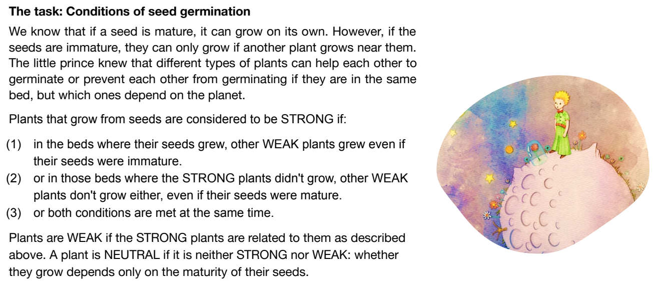
The task

The Little Prince explores the possibilities of rising plants on different planets.

He has seeds of three types of plants: X, Y and Z, which are visually distinguishable from each other. The seeds of each type are of two kinds: *mature* and *immature*, but they are not visually distinguishable from each other. The Little Prince's task is to see if plants of one type can help plants of another type to grow if their seeds are planted side by side.

For experiments, the Little Prince has several planets on which there are several types of soil: stony or sandy. On each planet there are 16 beds, in each of which only 3 seeds of plants can be planted, and all of them must be of different types: one of type X, one of type Y, one of type Z.

Figure 8. Task instructions page 1



The task: Conditions of seed germination

We know that if a seed is mature, it can grow on its own. However, if the seeds are immature, they can only grow if another plant grows near them. The little prince knew that different types of plants can help each other to germinate or prevent each other from germinating if they are in the same bed, but which ones depend on the planet.

Plants that grow from seeds are considered to be STRONG if:

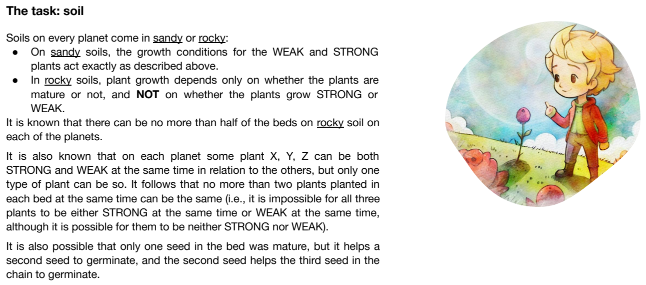
(1) in the beds where their seeds grew, other WEAK plants grew even if their seeds were immature.

(2) or in those beds where the STRONG plants didn't grow, other WEAK plants don't grow either, even if their seeds were mature.

(3) or both conditions are met at the same time.

Plants are WEAK if the STRONG plants are related to them as described above. A plant is NEUTRAL if it is neither STRONG nor WEAK: whether they grow depends only on the maturity of their seeds.

Figure 9. Task instructions page 2



The task: soil

Soils on every planet come in sandy or rocky:

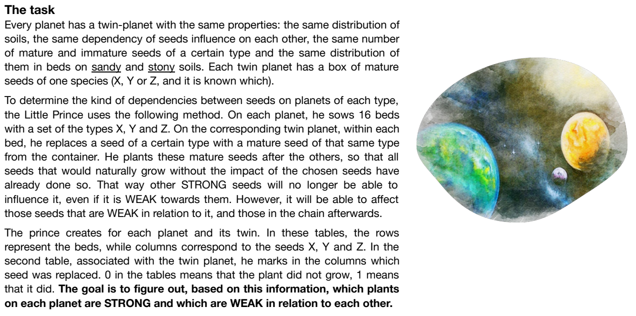
- On sandy soils, the growth conditions for the WEAK and STRONG plants act exactly as described above.
- In rocky soils, plant growth depends only on whether the plants are mature or not, and **NOT** on whether the plants grow STRONG or WEAK.

It is known that there can be no more than half of the beds on rocky soil on each of the planets.

It is also known that on each planet some plant X, Y, Z can be both STRONG and WEAK at the same time in relation to the others, but only one type of plant can be so. It follows that no more than two plants planted in each bed at the same time can be the same (i.e., it is impossible for all three plants to be either STRONG at the same time or WEAK at the same time, although it is possible for them to be neither STRONG nor WEAK).

It is also possible that only one seed in the bed was mature, but it helps a second seed to germinate, and the second seed helps the third seed in the chain to germinate.

Figure 10. Task instructions page 3



The task

Every planet has a twin-planet with the same properties: the same distribution of soils, the same dependency of seeds influence on each other, the same number of mature and immature seeds of a certain type and the same distribution of them in beds on sandy and stony soils. Each twin planet has a box of mature seeds of one species (X, Y or Z, and it is known which).

To determine the kind of dependencies between seeds on planets of each type, the Little Prince uses the following method. On each planet, he sows 16 beds with a set of the types X, Y and Z. On the corresponding twin planet, within each bed, he replaces a seed of a certain type with a mature seed of that same type from the container. He plants these mature seeds after the others, so that all seeds that would naturally grow without the impact of the chosen seeds have already done so. That way other STRONG seeds will no longer be able to influence it, even if it is WEAK towards them. However, it will be able to affect those seeds that are WEAK in relation to it, and those in the chain afterwards.

The prince creates for each planet and its twin. In these tables, the rows represent the beds, while columns correspond to the seeds X, Y and Z. In the second table, associated with the twin planet, he marks in the columns which seed was replaced. 0 in the tables means that the plant did not grow, 1 means that it did. **The goal is to figure out, based on this information, which plants on each planet are STRONG and which are WEAK in relation to each other.**

Figure 11. Task instructions page 4