Causal learning from actual cause judgements — Pre-registration

Nicolas Navarre
Supervision: Salvador Mascarenhas and Can Konuk
Institut Jean Nicod — École Normale Supérieure
Paris, France

February 3, 2023

Potential reviewers: Tom Icard, Chris Lucas, Neil Bramley, Stefano Palminteri.

Introduction

Background

Cognitive scientists are interested in addressing the question of how causal information is learned. Causal learning in psychology has been extensively studied in both adults and in children (Gopnik et al., 2004; Gopnik & Schulz, 2007). For Pearl, causal inference requires intervention such that causal relations can be disambiguated between correlations, which sits at layer-2 of Pearl's causal hierarchy (Pearl, 2009). Pearl's formalisms of causal structures were aimed to be used as a tool in computer science to enrich statistical inference, but have consequently become of great interest to cognitive scientists as a model of learning and reasoning.
While causal learning may describe the process by which general causes are learned, there have been results demonstrating a cognitive process with the notion of *actual cause*. *Actual cause* seems to distinguish between events in order to determine one as the *actual* reason for the outcome. This cognitive judgement seems to be mediated by two factors (Icard, Kominsky, & Knobe, 2017):

1. Causal structure of events
2. Normality of observable causes

Causal heuristics

The main finding is a three-way connection between the priors of the events, the existing causal structure and the actual cause judgements. You need information about two of the three in order to create the third. The direction which will be of interest in this project is identifying causal structure from normality priors and actual cause judgements. This can be done by leveraging some heuristics that can be derived from the actual cause judgements. Icard et al. (2017) manage to identify two strong heuristics that can be extracted from a two variable one outcome paradigm.

Abnormal inflation The canonical example of this notion is the fire that is caused by a lightning strike in a forest in the presence of oxygen. It is clear that both the oxygen and the lightning are general causes of a fire occuring when the lightning strikes a tree. However, there is a stronger cognitive judgement that points to the lightning rod as being the *actual cause*.
In this case the oxygen is so much more abundant (common) than the occurrence of a lightning strike. Since both are sufficient for the fire to occur, the *actual cause* is attributed to the less normal outcome. This heuristic is known as **abnormal inflation**.

Abnormal deflation If we consider however, a disjunctive outcome we see the opposite effect where the normal object is observed as the actual cause.

Suppose that a project needs to be approved by one of two bureaucratic entities. B1 is known to be very picky about approving projects, while B2 is known to approve most projects it receives. When you submit your project, you receive an approval from both parties. While both are known to be general causes of the approval, the normal approval (B2) remains as the *actual cause*. This phenomenon is referred to as **abnormal deflation**.

Models of Actual Cause

As of now, the **only** models of actual cause are those of (Icard et al., 2017; Quillien & Lucas, 2022), which seem to depend on at least some form of counterfactual simulations which may connect to Pearl's layer-3 in the causal hierarchy. It is clear that these kinds of judgements must play a role in causal learning, and have become of great interest to the field to identify why we may possess the kinds of cognitive processes that bring about these judgements. Some have appealed to the need for counterfactual simulations in order for optimal interventions in causal learning (Icard et al., 2017; Lee & Bareinboim, 2018).

Causal Rules

Causal rules are a type of causal structure that essentially amount to a logical operation based on the presence or absence of general causes of an outcome. For these reasons, we may reduce the causal structure to a deterministic causal rule or logical operation. The two canonical cases involve a rule of two events under disjunction and conjunction.

$$O \leftarrow A \land B \tag{1}$$

$$O \leftarrow A \lor B \tag{2}$$

While the actual cause heuristics involve two variables, we can imagine extending these cases to multiple events that as a group are preferentially considered as the plural cause as opposed to any singular event (Konuk et al. in progress). Given this, we can generalize the set of causal rules to be any combination of the observable events on the given outcome.

$$O \leftarrow (A \lor B) \land C$$
$$O \leftarrow A \land (B \lor C)$$
$$O \leftarrow (A \land B) \lor (C \land D)$$

Learning causal rules

A causal learner must make an inference on what events $A, B, C, \ldots$ will have an effect on some outcome O. One possible way of establishing the priors is by interacting or observing the data in the environment. If we take all observable events and their possible instances, we can tabulate them with their respective outcomes as in Table 1 (a). In order to enrich the data with actual cause information, the data can be augmented with some highlights that indicate the actual cause of the outcomes as in 1 (b).

Sample	A	B	O
1	🟩	🟥	🟥
2	🟩	🟩	🟩

(a) Data without actual cause judgements.

Sample	A	B	O
1	🟩	▌🟥	▌🟥
2	▌🟩▐	🟩	🟩

(b) Data with acutal cause judgement

Table 1: Tabular data examples for $O \leftarrow A \land B$, with B more normal than A.

We can see that the number of possible data inputs is 2^N, where N is the number of observables in the causal system. Furthermore, there are 2^{2^N} distinct causal rules that could lead to different outcome results. The goal of the causal learner is to learn what the rule is the would correspond to the observations provided.

Evaluation of learning

Instead of trying to determine what rule was learned by the causal learner, we can opt to identify their success rate on a set of data. Given that only 2^N distinct data are possible, this forces the design to split the data into an observational and testing phase. In this experimental design, we should opt for having fewer observational trials and more test trials. This methodological choice will allow for a finer grained range of learning results. The trade-off however, is that the fewer data used in the observation phase, the less variability we have in setting up the priors. This will be discussed in the methods section for the causal learning experiment.

The ideal learner

If we consider an ideal observer that makes optimal inference on the information received, we can formalize the amount of information encoded in the datum. In information-theoretic terms, each datum reduces the space of possible rules (of the 2^{2^N}) in half, until all 2^N data are shown, only one rule remains.

However, we might also want to consider how the actual cause judgements also constrain the set of possible rules. This will be a function of the abnormal inflation/deflation heuristics presented before, as well as potential plural cause heuristics. Of course, we do not expect people to correspond to the ideal reasoner, however it would provide a normative account of the extent that causal judgements ought to help in causal learning.

Research project

The aim of this study is to provide further evidence for the need for these judgements of *actual cause*. The focus of this project is to identify how a causal learner can learn the causal structure when these judgements of actual cause are provided to them as opposed to without. This research program should capture the informational value of communicating actual cause judgements, since the causal judgements come from another agent who presumably does know the causal structure.

Key research question: *How does the presence of an actual cause judgement, with the aim of learning a causal structure assist in the learning of the causal structure?*

General hypothesis

We expect people to perform better at causal learning tasks when they have actual cause judgements available to them. Furthermore, if this hypothesis is correct, we should expect the ”effectiveness” of the judgements in causal learning to depend on the following items.

- Diversity of information available
- Actual cause heuristics

Concrete Hypothesis The priors of the environment variables dictate the diversity of events that tend to happen in real life. If the priors are strongly skewed, it renders the variance of events in the environment very small. On the other hand, more even priors will lead to a larger distribution of kinds of events. In the case where the variance of events is small, we should expect the causal judgements to give a strong indication towards determining the causal structure when an abnormal outcome is observed. Conversely, when the variance is quite large, we can expect the judgements to play a smaller role, and the causal structure may be determined with the larger sample variance of events. Moreover, when actual cause judgments are available, their in formativeness will depend on the heuristics derived from the observations.

Methods – Causal learning experiment

The causal learning experiment will involve a 2x2 factorial design with the testing score as the dependent variables, and the following independent variables:

- Actual cause judgement provided
- Strength of priors on observable variables

This experimental design aims to capture the three-way interaction of the causal structure, priors on general causes and actual cause judgements. By having the participants in the condition with no actual cause judgements repeat the trial, only with actual cause judgements the second time, we will be able to extract both between-participant and within-participant effects of the presence of actual cause judgements. The strength of priors will remain exclusively between priors however.

Participants

We will recruit 160 participants on Prolific with 40 participants per condition. This is in accodance with the number of participants recruited for the actual casuse judgements in Quillien and Lucas's (2022) experiment. We will exclude participants with a background consisting in more than one graduate course in natural language semantics or pragmatics. We will exclude participants who fail to interpret the 2-gene data examples correctly. We will also exclude participants who fail to demonstrate the causal heuristics of abnormal inflation and abnormal disjunction in the 2-gene data examples. We will exclude participants who have any background knowledge on genomics and medical drug treatments. We will also exclude participants who reported using lots of notes or diagrams during the task.

Procedure

Experimental context

In the experimental setting, we will need to present an environment in which each observable and outcome has some meaning. This is typically hard to do without introducing inherent biases from pre-existing concepts about the observables. For this reason, we will introduce an experimental paradigm of the following form:

> *You are a medical student studying the outcome of different drugs used to treat headaches. Any patient who takes a drug will react either positively (no headache) or negatively (headache remains). The medical textbooks explain that the outcome of the drug's effect depends on the genetic codes of the patients. Suppose we have a database of patients (with their genetic data) who have taken and responded to a drug treatment either positively or negatively. This database also contains annotations indicating which genomes were* ***responsible*** *for the treatment outcome. Your job is to use this information to determine if new patients, knowing only their genetic data, will respond positively or negatively to the drug treatment.*

This poses two advantages for the experimental design. First, it allows us to assign any arbitrary rule to a kind of drug treatment without further need of an explanation for the causal rule. Second, it leaves the learning context devoid of existing priors from general purpose knowledge about genetics and drug responses.

Since our experimental design allows for both within-participant and between-participant effects of the actual cause judgements, we can leave in the information about the responsibility of each genome for the treatment outcome i.e. the actual cause judgments.

Verification of understanding

Prior to showing the causal learning condition, we want to make sure that the participants are capable of interpreting the data and their causal outcomes. We will start by showing them a table of patients' genetic data with the drug treatment outcomes as in (Table 1). Additionally, we would like to probe their ability to make use of the two main causal judgement heuristics of abnormal inflation/deflation. In order to achieve this, we will showcase the two canonical examples with two observable causes as described by rules (1) and (2). First, we will ask if the data match the known drug treatment. Next, we will ask the participants to annotate each of the data samples with each gene they think was the cause of the outcome in each trial. We will use these examples to ensure that the participants understand how to interpret the data as well as provide us with necessary judgements for the exclusion criteria.

Experimental Stimuli

After the primary screening, we will present the participants with their respective data tables of drug treatment outcomes for some unknown drug. Since the experimental task will involve 4 genes this means that we have 16 possible data samples to split between observations and tests. To get the best measure for their prediction scores we should opt for as many test items as possible. However, we must leave room for some diversity in the observations. By selecting 4 of the 16 possible observations this would allow room for us to modify the priors of the information by ”padding” the data with repeated samples. We will fill a table of 20 samples with 4 distinct genetic combinations that are repeated as necessary to adjust the priors on each genome’s occurrence. All four experimental contitions can be seen in Tables 2 and 3, which divide the conditions based on the strength of priors and presence of actual cause judgements.

Patient	A	B	C	D	O
1	🟩	🟩	🟥	🟥	🟩
2	🟩	🟥	🟩	🟩	🟥
3	🟩	🟥	🟩	🟩	🟩
4	🟩	🟩	🟥	🟥	🟩
5	🟩	🟩	🟥	🟥	🟩
6	🟩	🟩	🟥	🟥	🟩
7	🟩	🟩	🟥	🟥	🟩
8	🟩	🟥	🟥	🟩	🟥
9	🟩	🟥	🟥	🟩	🟥
10	🟩	🟩	🟥	🟥	🟩
11	🟩	🟩	🟥	🟥	🟩
12	🟩	🟩	🟥	🟥	🟩
13	🟥	🟩	🟩	🟥	🟩
14	🟩	🟥	🟥	🟩	🟥
15	🟩	🟩	🟥	🟥	🟩
16	🟩	🟩	🟥	🟥	🟩
17	🟥	🟩	🟩	🟥	🟩
18	🟩	🟩	🟥	🟥	🟩
19	🟩	🟥	🟩	🟩	🟩
20	🟩	🟥	🟥	🟩	🟥

(a) Data without actual cause judgements.

Patient	A	B	C	D	O
1	▌🟩	▌🟩	▌🟥	🟥	🟩
2	🟩	🟥	▌🟥	▌🟩	🟥
3	🟩	🟥	▌🟩	▌🟩	🟩
4	▌🟩	▌🟩	▌🟥	🟥	🟩
5	▌🟩	▌🟩	▌🟥	🟥	🟩
6	▌🟩	▌🟩	▌🟥	🟥	🟩
7	▌🟩	▌🟩	▌🟥	🟥	🟩
8	🟩	🟥	▌🟥	▌🟩	🟥
9	🟩	🟥	▌🟥	▌🟩	🟥
10	🟩	🟥	▌🟩	▌🟩	🟩
11	🟩	🟥	▌🟩	▌🟩	🟩
12	🟩	🟥	▌🟩	▌🟩	🟩
13	🟥	🟩	▌🟩	▌🟥	🟩
14	🟩	🟥	▌🟥	▌🟩	🟥
15	▌🟩	▌🟩	▌🟥	🟥	🟩
16	▌🟩	▌🟩	▌🟥	🟥	🟩
17	🟥	🟩	▌🟩	▌🟥	🟩
18	▌🟩	▌🟩	▌🟥	🟥	🟩
19	🟩	🟥	▌🟩	▌🟩	🟩
20	🟩	🟥	▌🟥	▌🟩	🟥

(b) Data with acutal cause judgement

Table 2: Stimuli for strong more even priors conditions $O \leftarrow (A \land B) \lor C$.

Patient	A	B	C	D	O
1	🟩	🟥	🟥	🟩	🟥
2	🟩	🟥	🟥	🟩	🟥
3	🟩	🟩	🟥	🟥	🟩
4	🟩	🟥	🟥	🟩	🟥
5	🟥	🟩	🟩	🟥	🟩
6	🟩	🟥	🟥	🟩	🟥
7	🟥	🟩	🟩	🟥	🟩
8	🟩	🟥	🟩	🟩	🟩
9	🟩	🟥	🟩	🟩	🟩
10	🟩	🟩	🟥	🟥	🟩
11	🟥	🟩	🟩	🟥	🟩
12	🟩	🟥	🟩	🟩	🟩
13	🟩	🟩	🟥	🟥	🟩
14	🟩	🟥	🟩	🟩	🟩
15	🟩	🟥	🟩	🟩	🟩
16	🟩	🟥	🟩	🟩	🟩
17	🟥	🟩	🟩	🟥	🟩
18	🟩	🟥	🟥	🟩	🟥
19	🟥	🟩	🟩	🟥	🟩
20	🟩	🟩	🟥	🟥	🟩

(a) Data without actual cause judgements.

Patient	A	B	C	D	O
1	🟩	🟥	🟥	🟩	🟥
2	🟩	🟥	🟥	🟩	🟥
3	🟩	🟩	🟥	🟥	🟩
4	🟩	🟥	🟥	🟩	🟥
5	🟥	🟩	🟩	🟥	🟩
6	🟩	🟥	🟥	🟩	🟥
7	🟥	🟩	🟩	🟥	🟩
8	🟩	🟥	🟩	🟩	🟩
9	🟩	🟥	🟩	🟩	🟩
10	🟩	🟩	🟥	🟥	🟩
11	🟥	🟩	🟩	🟥	🟩
12	🟩	🟥	🟩	🟩	🟩
13	🟩	🟩	🟥	🟥	🟩
14	🟩	🟥	🟩	🟩	🟩
15	🟩	🟥	🟩	🟩	🟩
16	🟩	🟥	🟩	🟩	🟩
17	🟥	🟩	🟩	🟥	🟩
18	🟩	🟥	🟥	🟩	🟥
19	🟥	🟩	🟩	🟥	🟩
20	🟩	🟩	🟥	🟥	🟩

(b) Data with acutal cause judgement

Table 3: Stimuli for the more even priors conditions $O \leftarrow (A \wedge B) \vee C$.

Measures

As briefly stated in the ”Evaluation of learning” section, we will be asking the participants to complete data tables with an empty column of outcomes. Since 4 of the possible data inputs will have been used for the observational data, this leaves 12 testable pieces of data. The quality of the learning will depend on how much better than chance the participants are at predicting the causal rule.

Predictions and analyses

We suspect that there will be an effect between the actual cause judgments and the strength of the priors. The effect will be characterized by an increase in prediction scores when judgments are provided, but only when the priors are strongly skewed. We also expect the prediction scores to be roughly better than chance in any scenario–assuming the data samples restrict the realm of possible rules to some extent. However, if some scores are systematically **worse** than chance, then that would be very strong evidence to support that certain *misleading* heuristics are being used based on the data provided prior to testing.

We suspect that this effect should generalize to most rules, however we do leave open the possibility that certain rules may be less accessible to use causal heuristics provided by actual cause judgements than others. To capture the full effect of this phenomena, we would have to explore a different kind of experimental design, which we will leave for future exploration.

Further considerations

Demonstrating priors On top of providing data for the observables and outcomes, it is important for the learners to have a good representation of the priors, especially in the use of actual cause heuristics.

There are several ways of presenting this information:

- Simply stating priors for each observable variable.
- Presenting the prior information as in a visual form i.e. bar chart, pie chart.
- Presenting repeated samples of the datum such that the total collection of samples represent the priors.

Learning and testing split It is still open to determine how many of the total rules should be sufficient to leave for the test phase. In the most extreme case, only one datum is provided to which a rule ought to be generalized. This would provide the strongest gradient of possible outcomes, however it does deviate from the feeling of evidence-based learning.

Biases towards learned rules Deviating from the ideal reasoner, it is reasonable to suspect that the learners will not consider every kind of rule evenly. We should suspect that certain rules are more attractive which might influence the effectiveness of the data.

Expected contributions

Salvador Mascarenhas (SM), Can Konuk (CK), Nicolas Navarre (NN)

- definition of research question — SM, CK, NN
- bibliography — CK, NN
- experimental design – SM, CK, NN
- programming experiment — NN
- creation and proofreading of the stimuli — CK, NN
- recruiting of participants — NN
- data analyses — SM, CK, NN
- data interpretation SM, CK, NN
- report writing — NN
- supervision of writing — SM, CK

References

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and bayes nets. *Psychological Review*, *111*(1), 3–32. Retrieved from https://doi.org/10.1037/0033-295x.111.1.3 doi: 10.1037/0033-295x.111.1.3

Gopnik, A., & Schulz, L. (Eds.). (2007). *Causal learning*. Oxford University PressNew York. Retrieved from https://doi.org/10.1093/acprof:oso/9780195176803.001.0001 doi: 10.1093/acprof:oso/9780195176803.001.0001

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80-93. doi: https://doi.org/10.1016/j.cognition.2017.01.010

Lee, S., & Bareinboim, E. (2018). Structural causal bandits: Where to intervene? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 31). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2018/file/c0a271bc0ecb776a094786474322cb82-Paper.pdf

Pearl, J. (2009). *Causality: Models, reasoning and inference* (2nd ed.). Cambridge University Press.

Quillien, T., & Lucas, C. G. (2022, June). Counterfactuals and the logic of causal selection.