Functional Rule Inference from Causal Selection Explanations

Nicolas Navarre[†] **(n.s.navarre@sms.ed.ac.uk)**
School of Informatics, School of Philosophy, Psychology & Language Sciences, University of Edinburgh
10 Crichton St, Edinburgh EH8 9AB

Can Konuk[†] **(can.konuk@ens.fr)**
Department of Cognitive Studies, Institut Jean-Nicod, Ecole Normale Supérieure
29 due d'Ulm, 75005 Paris, France

Neil R. Bramley (neil.bramley@ed.ac.uk)
Psychology Department, University of Edinburgh
7 George Square, Edinburgh, EH8 9JZ

Salvador Mascarenhas (salvador.mascarenhas@ens.fr)
Department of Cognitive Studies, Institut Jean-Nicod, Ecole Normale Supérieure
29 due d'Ulm, 75005 Paris, France

Abstract

Building on counterfactual theories of causal-selection, according to which humans intuitively evaluate the causal responsibility of events, we developed an experimental paradigm to examine the effect of causal-selection explanations on abductive causal inference. In our experiment, participants attempted to infer the rule responsible for winning outcomes of random draws from urns with varying sampling probabilities. Participants who were provided with causal-selection judgments as explanations for the outcomes made significantly closer inferences to the rule than those relying on observations alone, or on other explanations of causal relevance. We mirror these empirical results with a computational model of inference from explanation leveraging the theories of causal selection.

Keywords: causal inference, causal selection, counterfactual theories of causation

Introduction

Humans form elaborate causal models of the world, which allow them to understand, forecast, and influence the events around them. A central puzzle in cognitive science concerns how these causal beliefs are learned. Extracting causal conclusions from observations of events is a notoriously hard problem (Bareinboim et al., 2022; Bramley et al., 2015), and everyday inference settings often provide few opportunities to perform the *interventions* (or experiments) needed to reliably infer the causal structure behind a distribution of events.

To mitigate these limitations, it seems plausible that people should frequently rely on social learning, to piggyback on the causal knowledge of one another in order to achieve an understanding of the world more efficiently. This lines up with the ubiquity of *causal explanations* in everyday discourse. From infancy, we frequently ask our peers for explanations for "why" things occur the way they do. As an everyday example, one might ask a neighbor 'Why did your flowers grow so well?' and learn something new from the explanation one receives (e.g. 'Because I used fertilizer').

A complicating feature of such everyday causal discourse is that explanations rarely lay out a complete mechanism sufficient to reproduce the explanandum, as we might expect from scientific textbook explanation. Rather, they tend to highlight one or a few of the causal factors involved and claim these as *the cause* of the event. You might point to fertilizer as the cause for the growth of these flowers, *rather than* for example the presence of the sun or water. This explanation seems reasonable in spite of the knowledge that sun and water are also prerequisites for flowers to grow, and would certainly have their place in an exhaustive causal theory of flower growth. Judgments of this kind, which single out a particular subset of causal variables as holding particular importance, are known in the psychological literature as *causal selection* (Quillien & Lucas, 2023), or causal responsibility judgments (Lagnado et al., 2013).

On the face of it, such selective explanations may appear to be poor conveyors of causal knowledge: by singling out a subset of variables in a system that often contains many more interrelated parts, they run the risk of reflecting only the explainer's preference for one kind of explanation. As such they might impoverish, rather than enrich, a requester's causal understanding. Yet, the psychological literature on causal selection has shown that people hold very consistent intuitions as to which of several events in a causal model is the most important cause of an outcome (Morris et al., 2018). This seems to unlock the possibility that people reverse engineer aspects of an explainer's beliefs about a causal system from the causal factors they choose to highlight in their explanations of specific outcomes.

We propose a novel experiment design to test this possibility. We put experimental subjects through a task of abductive causal inference, where they have to retrieve the causal structure underlying a dataset from a mixture of observational and explanatory evidence. Our design allows us to control the main known drivers of causal selection judgments. Our results show that causal-selection explanations help subjects generalize more adequately from limited data than when

[†]These authors contributed equally.

they are provided with observational data alone, or with other causal explanations that do not point at the main causes targeted by causal-selection judgments.

Theoretical Background

Recovering what structures and functional relationships underlie a system based on observations of the states of that system can be a particularly difficult task.

The crux of the challenge lies in the fact that all too often multiple distinct causal hypotheses might be equally compatible with a set of observations. Observing for example that an event E regularly follows the occurrence of two events A and B doesn't help me decide whether the underlying structure is one where the conjunction of A and B causes E to occur ($E \leftarrow A \wedge B$), of one where either one of A and B would have caused it to occur ($E \leftarrow A \vee B$) — here restricting the focus only to rules involving Boolean variables and connectives, for simplicity. A greater variety of observations might help narrow down the possibilities: if A occurs but B doesn't, while E still follows, I can exclude the possibility that A and B are both necessary for E to occur. I would still need additional observations however to rule out other possibilities, such as $E \leftarrow A \vee B$ or simply $E \leftarrow A$, or any other rule consistent with this limited set of observations.

The problem becomes all the more acute in settings where crucial observations occur infrequently. Whenever some of the events relevant to a causal system have very high or very low probabilities of occurrence, one rarely gets the chance to observe crucial event combinations. If A is always present, I cannot learn what the effect of its absence would be. Yet this is crucial information to infer A's causal relationship to E.

Remarkably enough, it is in those situations where the available observations are rather poor at covering the space of causal systems that causal-selection judgments are going to be particularly sharp, and exhibit strong preference in favor of certain variables.

Causal selection judgments

Causal selection judgments have been shown to be sensitive to two main factors: the causal *rule* that links candidate causes with the relevant outcome, on the one hand, and the *normality* associated with the different variables, on the other (Morris et al., 2019; Icard et al., 2017; Quillien & Lucas, 2023).

For example, in a situation where several different variables are each individually *necessary* for an outcome to occur (for example, when both water and fertilizer are required for my flowers to grow), subjects tend to think of the most unusual variables (the fertilizer) as 'the cause', and comparatively disregard the importance of the most expected ones (the water), a pattern of judgment known as *abnormal inflation*. By contrast, in a situation where each variable is individually *sufficient*, people tend to favor the most probable events as explanations. This latter pattern of judgment is known as *abnormal deflation* (Icard et al., 2017).

Counterfactual theories

Two successful theories of these patterns of causal judgment to date involve the notion of *counterfactual sampling* (Icard et al., 2017; Quillien & Lucas, 2023). According to counterfactual theories, causal-selection judgments involve a two-step process. First, one uses one's causal model to generate a sample of counterfactual situations, where the values of causal variables differ from what has actually occurred and the outcome potentially differs too. The frequency of each event across counterfactuals is a function of their prior probabilities, and whether or not the event effectively happened in the real world. The outcome of interest is determined by the events sampled, following subjects' causal models of the situation. Empirically, it seems that people consider counterfactuals that are both *likely* under the causal model of the situation, and *close to the observations they have made*. From a sample of counterfactuals one can compute a causal responsibility score, as some measure of the covariation between the states of causal variables and the outcome (different across models). The variables with the highest causal responsibility score are those that subjects are expected to favor in causal-selection judgments (Quillien, 2020).

Inference from explanations

Not only do subjects show a lot of consistency in these patterns of judgments, they are also eager to assume that others follow similar patterns of judgments, and derive pragmatic inferences out of such assumptions. As shown by Kirfel et al. (2022), in a situation where two causes A and B are known to impact an event E, subjects told that E happened 'because of A' will infer that the underlying causal structure is $E \leftarrow A \wedge B$ when A is the more expected variable, and $E \leftarrow A \vee B$ when A is the more unexpected variable (in a situation where they are to choose between just those two structures). More broadly, it has also been shown that certain types of explanation serve as a guide to property generalization for both children and adults (Lombrozo & Gwynne, 2014; Vasil et al., 2022).

Experiment

Here we present a novel experiment design, to show that this capacity to derive inferences from causal selection judgments can also help abductive causal inference in conditions closer to everyday life, where the causal structure to be guessed is of relative complexity, the space of possibilities open-ended, and the available observational data too limited to infer the rule with deductive certainty. This extends previous accounts of causal inference from explanation (Nam et al., 2023; Lampinen et al., 2021) by focusing on the role of causal-selection judgments specifically, rather than just considering the role of any causal explanations in inference.

Design

The game Participants are invited to participate in a game where they must infer a hidden rule based on examples of winning and losing outcomes. The game involves four urns, each containing a mix of colored and uncolored balls, with

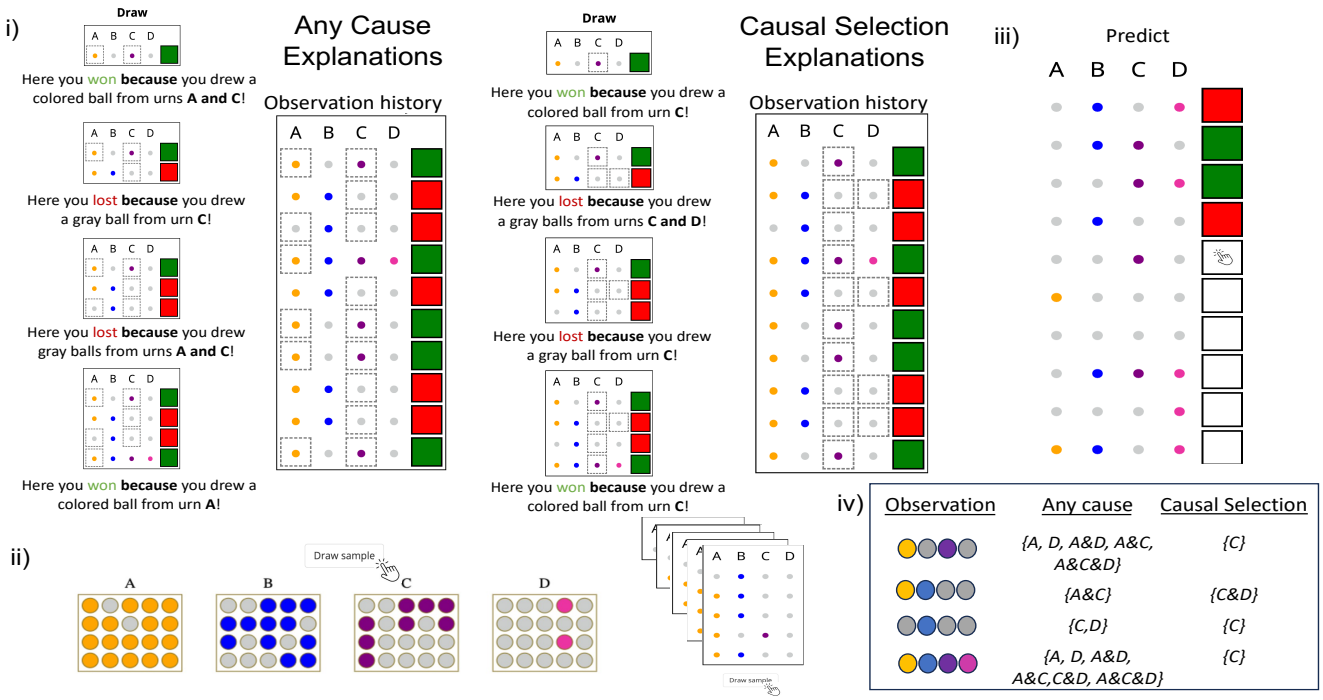

Figure 1: The experiment design; (i) shows a sequence of draws with all four samples in the experiment along with respective explanations. To the right of the samples is the entire history of observations once all 10 samples are drawn; (ii) shows the urns participants sample from and a visual example of the sampling process; (iii) shows the table of samples participants are tasked with predicting (not showing all 16). Participants have access to their observational history when making predictions. (iv) shows a comprehensive list of the observations and explanations given in the experiment.

each urn having a distinctive ball color, as in Figure 1-(ii). A round of the game involves drawing a ball at random from each of the four urns, with the result of these draws determining whether the player wins or loses the round. Whether a given draw from the urns corresponds to a win or a loss is determined by a fixed rule, but subjects are not told what the rule is. Their task in the experiment is to guess it.

The conditions The experiment involved three between-participants conditions, two of which are depicted and described in Figure 1-(i).

In one condition, not depicted in Figure 1-(i) for reasons of space, participants only have access to observational data: they get to draw several times from the four urns and observe the outcome of each draw. To illustrate, on one particular draw, they might draw a colored ball from urns A, and C, a white ball from urns B and D, and then observe that this particular draw corresponds to a win, as in the first row of Figure 1-(i), minus the dotted squares. Such an observation gives them some information about the rule that links draws to outcomes. For example, it tells them that drawing a colored ball from urn D is not necessary for one to win in this game. We call this condition *observation only* or OBS.

In the other two conditions, participants see the same observations, but on top of that, they also have access to some *explanations* which tell them, for each draw that they observe, *why* they won or lost in that particular round of the game. These explanations point to a subset of the balls drawn as being *responsible* for the outcome of the game. These are the dotted squares around balls drawn from urns in Figure 1-(i).

In the *causal selection* condition, or CS, participants are given explanations that correspond to intuitive causal-selection judgments that a person knowing the causal rule would have been expected to make. We chose the relevant judgments for each observation by computing the predictions of two models known to provide a good approximation to human causal-selection judgments (see details below on causal-selection explanations).

In the *any cause* condition, or AC, participants are also given causal explanations for each observation, but these do not match intuitive causal-selection judgments. Instead, they point to any subset of the variables featured in an observation that played an active causal role, *except* for the one subset of variables which our best theories of causal-selection judgments predict to be the most important causes. The rationale behind this condition was to make sure that causal-selection judgments did not help subjects just by virtue of the fact that they point to any variable that made a contribution to the outcome, which could have provided a first step towards reconstructing the causal rule.

Materials

The observations that subjects saw consisted of random draws from the four urns represented in Figure 1-(ii). Probability acted as a proxy for normality in our design. Each urn contained a different mixture of colored and white balls, indicating the following probabilities of drawing colored balls from urns: $P(A) = 0.9$, $P(B) = 0.6$, $P(C) = 0.4$, $P(D) = 0.1$. The position of the urns was randomized across subjects, but for ease of exposition we will refer to those urns by the names in Figure 1-(ii). The use of urns allowed us to have a direct handle on participants' subjective probabilities, a paradigm that has proved effective in past experiments on causal-selection judgments (Morris et al., 2018; Quillien & Lucas, 2023; Konuk et al., 2023).

The rule The rule that determined the outcome, across all conditions, was as follows. To win, one must either draw a colored ball from both the high-probability urn A and the low-probability urn D, or from the intermediate probability urn C. In logical notation, this corresponds to

$$\text{WIN} \leftarrow (A \wedge D) \vee C. \tag{1}$$

We chose this rule because its logical form involves both conjunction and disjunction, so that we expect causal-selection judgments to be sensitive to both the abnormal inflation and deflation effects, as well as complex combinations of the two, depending on the target observation.

The Observations In order to guess the rule in (1), subjects were provided with the 10 observations in Figure 1-(i). Participants drew observations successively from the urns by clicking the 'Draw sample' button above the urns. The order in which they appeared was randomized across subjects, but all participants saw functionally identical observations. Many of these observations were repeated, so that subjects only saw *four* unique observations in total, listed in Figure 1-(iv).

The choice of observations was constrained by four desiderata: (1) they had to be consistent with the probabilities implied by the urns, avoiding observations that the priors made too unlikely; the repetitions made sure that the frequency with which each color is drawn is proportional to its probability; (2) subjects had to draw a colored ball and a white ball from each urn at least once (to convey the idea that for each draw they observed, the alternative draw was a live possibility); (3) the two models of causal-selection judgments that we used as benchmarks had to agree as to the most important cause of the outcome for each observation (see next section for details); (4) for each observation, there had to be at least one active cause for the outcome (this mattered for the *any Cause* condition, see below).

Causal selection explanations. The causal strength of explanations presented in the CS and AC conditions were computed using two models of causal-selection judgments, the Counterfactual Effect Size Model (CESM; Quillien & Lucas, 2023) and the Necessity and Sufficiency Model (NSM; Icard et al., 2017).

We considered these two theories because they have been shown to provide good predictions of subjects' judgments in a variety of documented cases (Quillien & Lucas, 2023). However, our goal in this study was not to commit to one particular model of causal-selection judgments. Rather, we meant to probe whether explanations can help subjects in a causal-inference task without assuming a particular theory of how these explanations are generated.

In both theories, the causal responsibility of each event that influenced an outcome is a function of three main parameters: (1) the prior probabilities of drawing a colored ball from each urn, (2) the balls that were actually drawn in the case under consideration, (3) the causal rule that determines the outcome. They also include a sampling parameter s, which represents the extent to which the counterfactual worlds from which the causal impact of an event is computed are anchored to the actual world of reference. We reused the values of that parameter that had been previously fit to behavioral data (Quillien & Lucas, 2023), namely $s = 0.73$ for the CESM and $s = 0.15$ for the NSM.

Both models were run on each of the possible selections of variables. This includes the conjunctions of these individual candidate causes, (such as $A \wedge C$) or 'plural causes', as humans also hold consistent intuitions about such causal combinations, which are sensitive to the same factors as those driving judgments for singular events (Konuk et al., 2023).

Given causal scores computed in this way, we selected the subset of draws whose causal score was highest as the explanation to be given to participants as explanations in the CS condition. Both models agreed on each of the four observations as to which event had the highest causal responsibility. The highest scoring events were highlighted with grey dotted boxes as in the Causal Selection Explanations table of Figure 1-(i). The explanations were also delivered linguistically to subjects for each observation as they drew them, as illustrated in the same figure.

***Any cause* explanations.** In the AC condition, the explanations we gave to subjects were sampled at random from any of the active causes of the outcome except for the one that was selected for causal selection judgments (see the full list in Figure 1-(iv)). Once an explanation had been provided for a given observation, we kept the same explanation for every repetition of that observation that a subject drew. The explanations were displayed in exactly the same way as in the CS condition. We took advantage of the fact that the locution 'X because Y' is one that can be used both for causal selection and for more generic causal attributions (Copley, 2020).

Scoring inference After subjects see the full ten observations, they are asked to make predictions for each of the 16 possible draws from these four urns. They are presented with a full table of possible draws as the one in Figure 1-(iii), with the boxes corresponding to the outcomes left blank. They should click on the boxes to turn them green or red, depend-

Fixed Effects	log-odds ratio	Standard Error	p-value
Intercept (OBS)	1.11562	0.07304	$<$ 2e-16
AC	-0.29284	0.10178	0.00401
CS	0.31959	0.10707	0.00284
Random Effects	**Variance**	**Std. Deviation**	
Participant	0.1972	0.4441	

Table 1: Results of Mixed-Effects logistic regression: `Sample-Accuracy ~ 1 + Condition + (1| Participant)`

ing on what they think the outcome would be for each particular draw. We recorded the accuracy of each prediction made in this way, with subjects scoring 1 for a row if they gave a prediction matching what the outcome that the rule in (1) determines for that row, and 0 otherwise.

Procedure

We recruited 298 participants on Prolific from the United States, United Kingdom, and Canada. Each participant was randomly assigned to one of the 3 conditions (AC: 98, CS: 97, OBS: 103).

First we explained the mechanics of urn sampling and the relationship between the number of colored balls in an urn and the probability of drawing one. We had participants play a much simplified version of the game, involving just two urns, to illustrate how a rule mediated the relation between draws and outcomes, and the workings of the testing procedure that would follow. In the relevant conditions, we also gave them examples of explanations, making sure to pick examples were CS and AC couldn't differ, so as not to prime their interpretation of subsequent explanations one way or the other.

Participants were then invited to make ten "dry" draws from the urns, as in Figure 1-(ii) (right), where they weren't provided with any outcomes, so as to get them to internalize the probabilities associated to each urn. The draws were randomized so as to reflect the probabilities.

They then drew the ten observations with outcomes from Figure 1-(i), accompanied with explanations depending on the condition, and subsequently offered their predictions for all 16 possible samples. Finally, participants completed a brief questionnaire, where they had the opportunity to describe the rule they had in mind in prose if they so wished, and were asked some demographics questions, before being redirected to Prolific for payment. The experiment was programmed using the `JsPsych` JavaScript library (de Leeuw et al., 2023).

Results

CS explanations helped subjects reach more accurate generalizations, while AC explanations made them less accurate Participants' accuracy across all 16 samples is summarized in Figure 2. As suggested by the plot, subjects in the CS condition were overall more accurate than in either the OBS

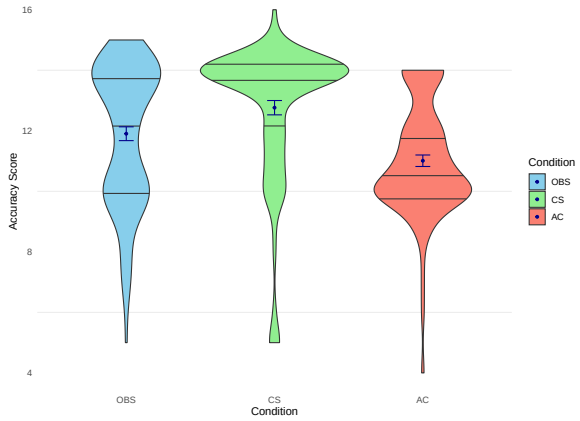

Figure 2: Prediction accuracy per condition. The blue dots represent the means per condition 12.76 (CS), 11.93 (OBS), 11.01 (AC). Error bars represent the standard error around the means. Solid black lines mark the medians and quartiles.

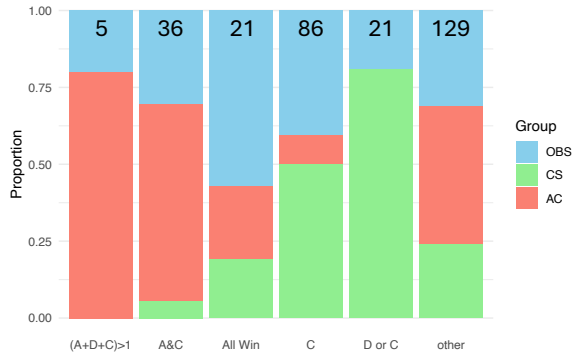

Figure 3: Most common rules inferred by the participants. The numbers in the bars represent the total number of participants in each group.

or AC conditions. A three-way ANOVA confirmed that the difference between the means of each condition was highly significant ($Df = 2, t = 15.65, p <$ 4e-07).

To further assess the effect of condition on subjects' accuracy, we ran a mixed-effects model using the accuracy of predictions for each row as dependent variable, the condition as fixed effect, and individual subjects as random effects. Results are summarized in Table 1. They confirmed the trend suggested by the figure: while CS taken as a factor had a positive effect on the correctness of guesses, AC had negative effect (compared to the OBS condition as baseline).

Removing the condition factor from the model resulted in a significantly worse fit to the data ($Df = 2, -\text{LogLik} = 2674.0$, BIC $= 5381.791$ for the full model; $Df = 4$, $-\text{LogLik} = 2690.1$, BIC $= 5397.057$ for the Intercept-only model; $\chi^2 = 32.205$; $p <$ 1.017e-07), confirming that the condition subjects were placed in significantly affected their generalizations in the expected direction.

Subjects converged towards certain high-scoring rules Because subjects gave an answer for all 16 observations possible with the available four urns, we were able to reconstruct the rule that guided their choices by looking at the truth table of their responses. Figure 3 plots the most popular patterns of responses per condition, translated into logical propositions that matched the contents of their responses. An outstanding pattern was the simple rule C, which was very popular in both the OBS and CS condition, although not in the AC condition.

Computational Modelling

To help interpret the results from the experiment, we constructed a Bayesian model that makes inferences about possible rules from the observations and explanations. The model is comprised of two elements:

i) A prior probability distribution over all rules one could represent with four urns. We only retained deterministic rules consisting of Boolean combinations of colored and white balls from each of the four urns, giving us a total of 2^{16} distinct rules up to equivalence. A simplicity prior was applied to this probability distribution, which penalized rules whose definition depended on a greater number of different variables (Lu et al., 2008; Lucas et al., 2015).

ii) A likelihood function to update one's probability as a function of new observations and explanations. Observations simply update the probability by excluding rules incompatible with a given observation O and renormalizing probabilities over the remaining rules. For explanations, the model first computes a causal score for every possible explanation $E_m \in \mathbf{E} = \{E_1, \ldots, E_n\}$ and rule $R_m \in \mathbf{R} = \{R_1, \ldots, R_{2^{16}}\}$, by taking the square of the causal responsibility score $\kappa(E_m, O, R_m)$ that E_m would get for O, under the assumption that the correct rule is R_m, using the CESM model. Then, the model uses that causal score as the basis for the likelihood $P(E \mid O, R)$ of each explanation by normalizing over the causal score of all possible explanations, following equation (2).

$$P(E \mid O,R) = \kappa(E,O,R)^2 \left(\sum_{E_i \in \mathbf{E}} \kappa(E_i,O,R)^2 \right)^{-1} \quad (2)$$

The posterior distribution over rules is then updated based on how well these likelihoods predict the explanation that the learning model sees in each condition.

Model Results and Analysis

We compare model results in each condition on three dimensions: i) The Maximum A Posteriori hypothesis (MAP): which rule has the highest posterior after the four distinct observations; (ii) the position of the intended rule $(A \wedge D) \vee C$ in the posteriors' rankings; (iii) the weighted score of each distribution, i.e. the score that each rule in $\mathbf{R}$ gets in our experiments' test, weighted by their posterior probability.

As shown in Table 2, CS explanations reliably rank the intended rule among the most probable candidate generalizations for the observed data, compared to the observation-only condition and the AC explanations on average. Additionally, the weighted score of the CS condition is significantly greater than that of both other conditions. Finally, it appears that the model accurately captures the attractiveness of the rule C, which stood out as the MAP in both the OBS and CS conditions, especially compared with the AC conditions, which favored a greater diversity of rules as MAP (with 5/60 AC configurations choosing C), in line with the distribution of that strategy across conditions, as reported above.

	OBS	**AC**	**CS**
Obs. 1	323	15684 ± 182	22
Obs. 2	43	5985 ± 52	6
Obs. 3	37	2577 ± 23	4
Obs. 4	67	952 ± 10	14
MAP rule	C	—	C
MAP score	14.00	10.6 ± 0.037	14.00
Weighted score	10.00	10.36 ± 0.013	**11.65**

Table 2: Upper section: the posterior probability ranking of the intended rule $(A \wedge D) \vee C$ after each observation. Below: the rule that has the highest posterior probability (MAP) after the fourth observation in each condition, and the weighted score of the respective final probability distributions.

Another takeaway from the model results is that, even in the CS condition, the intended rule didn't come out as the MAP. Later iterations of this design will address this by examining cases in which we can expect CS explanations to guarantee the exact ground-truth rule underlying the game assuming optimal inferential abilities. In any event, these results concur with our experimental results in that the CS explanations bring the intended rule as the MAP much more reliably in light of new observations than the other two conditions.

Discussion and Conclusions

Our findings indicate that causal selection judgments serve as valuable cues when inferring causal structures from limited observational data. Our experiment is the first to provide evidence of this in a context where the causal rule underlying the data is of relative complexity and the space of possible hypotheses open ended.

Individuals not only demonstrate improved generalization from the data when causal-selection judgments are provided as explanations, they also exhibit notably poorer performance when presented with true but less selective causal explanations. These findings, in conjunction with the results from our computational model, strongly suggest that causal-selection judgments can aptly tap into humans' shared set of intuitions about causality to convey elaborate causal knowledge via relatively simple explanations.

Acknowledgments

We thank Tadeg Quillien, Thomas Icard, and the LANG-REASON team at Ecole Normale Supérieure for invaluable feedback. This work was supported by *Agence Nationale de la Recherche* grant ANR-18-CE28-0008 (LANG-REASON; PI: Mascarenhas), UKRI grant EP/S022481/1 and EURIPANR-17-EURE-0012 (Investments d'Avenir Program), and by Ecole Doctorale Fronti'eres de l'Innovation en Recherche etEducation—Programme Bettencourt.

References

Bareinboim, E., Correa, J. D., Ibeling, D., & Icard, T. (2022). On Pearl's Hierarchy and the Foundations of Causal Inference. In *Probabilistic and causal inference: The works of judea pearl* (1st ed., p. 507–556). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3501714.3501743

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 708–731. Retrieved from http://dx.doi.org/10.1037/xlm0000061 doi: 10.1037/xlm0000061

Copley, B. (2020). Events are the source of causal readings in the simplest English conditionals. In S. Kaufmann & D. Over (Eds.), *Conditionals - Logic, Linguistics, and Psychology.* Retrieved from https://hal.science/hal-02431650

de Leeuw, J. R., Gilbert, R. A., & Luchterhandt, B. (2023). jspsych: Enabling an open-source collaborative ecosystem of behavioral experiments. *Journal of Open Source Software*, *8*(85), 5351. Retrieved from https://doi.org/10.21105/joss.05351 doi: 10.21105/joss.05351

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

Kirfel, L., Icard, T., & Gerstenberg, T. (2022, July). Inference from explanation. *Journal of Experimental Psychology: General*, *151*(7), 1481–1501. Retrieved from https://doi.org/10.1037/xge0001151 doi: 10.1037/xge0001151

Konuk, C., Goodale, M., Quillien, T., & Mascarenhas, S. (2023, May). *Plural causes in causal judgment.* PsyArXiv. Retrieved from psyarxiv.com/nuptb doi: 10.31234/osf.io/nuptb

Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013, July). Causal responsibility and counterfactuals. *Cognitive Science*, *37*(6), 1036–1073. Retrieved from http://dx.doi.org/10.1111/cogs.12054 doi: 10.1111/cogs.12054

Lampinen, A. K., Roy, N. A., Dasgupta, I., Chan, S. C. Y., Tam, A. C., McClelland, J. L., ... Hill, F. (2021). Tell me why! - explanations support learning of relational and causal structure. *CoRR*, *abs/2112.03753*. Retrieved from https://arxiv.org/abs/2112.03753

Lombrozo, T., & Gwynne, N. Z. (2014, September). Explanation and inference: mechanistic and functional explanations guide property generalization. *Frontiers in Human Neuroscience*, *8*. Retrieved from http://dx.doi.org/10.3389/fnhum.2014.00700 doi: 10.3389/fnhum.2014.00700

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*(4), 955–984. doi: 10.1037/a0013256

Lucas, C., Griffiths, T., Williams, J., & Kalish, M. (2015, 03). A rational model of function learning. *Psychonomic bulletin & review*, *22*. doi: 10.3758/s13423-015-0808-5

Morris, A., Phillips, J., Gerstenberg, T., & Cushman, F. (2019, August). Quantitative causal selection patterns in token causation. *PLOS ONE*, *14*(8), e0219704. Retrieved from http://dx.doi.org/10.1371/journal.pone.0219704 doi: 10.1371/journal.pone.0219704

Morris, A., Scott-Philips, J., Icard, T. F., Knobe, J., Gerstenberg, T., & Cushman, F. (2018). *Judgments of actual causation approximate the effectiveness of interventions.* (Psy ArXiv) doi: 10.31234/osf.io/nq53z.

Nam, A., Hughes, C., Icard, T., & Gerstenberg, T. (2023, May). Show and tell: Learning causal structures from observations and explanations. In *Proceedings of the 45th annual conference of the cognitive science society.* Center for Open Science. Retrieved from http://dx.doi.org/10.31234/osf.io/wjs9q doi: 10.31234/osf.io/wjs9q

Quillien, T. (2020). When do we think that x caused y? *Cognition.*, *205*. doi: 10.1016/j.cognition.2020.104410

Quillien, T., & Lucas, C. G. (2023). Counterfactuals and the logic of causal selection. *Psychological Review*. doi: 10.1037/rev0000428

Vasil, N., Ruggeri, A., & Lombrozo, T. (2022, January). When and how children use explanations to guide generalizations. *Cognitive Development*, *61*, 101144. Retrieved from http://dx.doi.org/10.1016/j.cogdev.2021.101144 doi: 10.1016/j.cogdev.2021.101144