UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
Paradoxical parsimony: How latent complexity favors theory simplicity

Permalink
https://escholarship.org/uc/item/33r065cs

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0)

Authors
Gong, Tianwei
Valentin, Simon
Lucas, Chris
et al.

Publication Date
2024

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Paradoxical parsimony: How latent complexity favors theory simplicity

Tianwei Gong (tia.gong@ed.ac.uk)[1], **Simon Valentin (s.valentin@ed.ac.uk)**[2],
Christopher G. Lucas (clucas2@inf.ed.ac.uk)[2], **Neil R. Bramley (neil.bramley@ed.ac.uk)**[1]
[1]Department of Psychology, University of Edinburgh, Edinburgh, Scotland, United Kingdom
[2]School of Informatics, University of Edinburgh, Edinburgh, Scotland, United Kingdom

Abstract

Investigating how people evaluate more or less complex causal theories has been a focal point of research. However, previous studies have either focused on token-level causation or restricted themselves to very small sets of explanatory variables. We provide a new approach for modeling theory selection that foregrounds the balance between observed and latent structure in the mechanism being explained. We combine a Bayesian framework with program induction, allowing an unbounded and partially observable model space through sampling, and reflecting how a preference for simplicity emerges naturally in this setting. Through simulation, we identify two rational principles: (1) Simpler explanations should be favored as latent uncertainty (the number of hidden variables) increases; (2) latent structure is attributed a larger role when the observable patterns become less compressible. We conducted a behavioral experiment and found that human judgments tended to reflect these principles, indicating that people are sensitive to latent uncertainty when selecting between explanations.

Keywords: explanation; mechanism; program induction; simplicity; hidden variables; inductive reasoning

Introduction

Human beings acquire causal theories by interacting with their environment and observing the factors that influence outcomes. These theories are then propagated through social interactions, as people share their knowledge by providing one another with causal explanations. Typically, these explanations fall short of capturing their explananda perfectly, leaving unexplained variance and uncertain generalizability. Even in cases where an explanation offers an almost perfect account of the available data, people might prefer a simpler alternative. This raises deep questions: What is the right level of complexity for a theory? And, what does this balance depend on?

Investigating how people evaluate more or less complex explanations of the same evidence has been a focus of past research (Lombrozo, 2007; Zemla, Sloman, Bechlivanidis, & Lagnado, 2023; Johnson, Valenti, & Keil, 2019). However, there are several typical features in previous studies that limit their generality. Firstly, many studies (Lombrozo, 2010; Pacer & Lombrozo, 2017; Johnson et al., 2019) asking explanatory complexity concentrate on the diagnostic or *token* level of explanation (i.e. a particular explanation for how or why a specific event happened), fewer have concentrated on the ways people make predictions or represent mechanisms at the *type* level of causation (i.e. a general theory to explain how a type of events comes into being). In particular, we lack a computational theory for determining when a type-level explanation *should* be simple or complex, given the evidence that has been observed so far.

Secondly, studies often ask participants to express or evaluate explanations written in natural language (Zemla et al., 2023; Sulik, van Paridon, & Lupyan, 2023). However, this might prompt participants to think from a communication perspective, which could elicit explanations that are distinct from communicators' representation or understanding (Lombrozo, 2010).

Thirdly, studies have predominantly focused on a few specific causal structures, such as the common-effect (e.g. where a symptom might stem from one disease or several; Pacer & Lombrozo, 2017; Johnson et al., 2019) or chain structures (e.g. where intermediary variables are incorporated into an explanatory narrative or sequence of events; Johnson & Ahn, 2015; Johnson et al., 2019). They also predominantly focused on a limited set of relationships (mainly generative) or functional forms, often noisy-OR disjunctive combinations of causes (Bramley, Lagnado, & Speekenbrink, 2015; Bramley, Dayan, Griffiths, & Lagnado, 2017; Gong, Gerstenberg, Mayrhofer, & Bramley, 2023; Kushnir, Gopnik, Lucas, & Schulz, 2010). This is largely a consequence of computational convenience since modeling inferences over a larger hypothesis space that covers a variety of functions as well as connectivity patterns rapidly gets unwieldy, making it difficult or intractable to compute the posterior distributions necessitated by a traditional Bayesian analysis.

In this study, we will focus on type-level explanations, deviating from analyzing natural language explanatory narratives, and taking an approach that supports studying induction in an unbounded hypothesis space. One way to make sense of inference in an unbounded hypothesis space is adopt a program induction approach — assuming learners generate explanatory hypotheses by composing them stochastically from a sufficiently expressive grammar. We here consider probabilistic context-free grammars (PCFGs), and concretely assume a grammar that can be used to produce any causal rule expressible in propositional logic applied to the question of how a set of putative causes combine to determine an effect (Buchanan, Tenenbaum, & Sobel, 2010; Griffiths & Tenenbaum, 2007). Similar formalisms have been applied to a range of concept learning settings (Bramley & Xu, 2023; Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Zhao, Lucas, & Bramley, 2024; Fränken, Theodoropoulos, & Bramley, 2022), and recent work has argued for the framework's general applicability to explaining our capacity for induction

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.* ©2024 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

(Piantadosi, 2021; Bramley, Zhao, Quillien, & Lucas, 2023). A core feature of a PCFGs is a built-in inductive bias favoring shorter and simpler explanations – longer strings typically involve more rule applications, which typically imply geometrically decreasing probabilities.

Using this representational approach, this paper will focus on a specific theoretical question: How do beliefs about the existence of hidden components in the environment influence the selection between more complex vs. simple causal explanations? A familiar domain, such as a physical or social setting, could contain different levels of complexity due to the number of hidden variables (Johnson et al., 2019; Lucas, Holstein, & Kemp, 2014; Valentin, Bramley, & Lucas, 2022; Rottman, Ahn, & Luhmann, 2011). Research has found that people spontaneously infer hidden causes in their environments (see Rottman et al., 2011, for a review). Models that include assumptions about these hidden causes have been shown to better explain human causal judgments than models that ignore them (Cheng, 1997; Luhmann & Ahn, 2007). From a normative standpoint, the states of hidden variables are unknown. However, knowing that these hidden variables exist and could feature in the mechanism of interest should influence how we explain the roles played by the explicit variables. In particular, this affects how much variance or noise we should tolerate in our explanations.

We used a causal learning paradigm, where the learner would judge what caused an effect given a set of evidence. As shown in Figure 1, the evidence can be summarized as logic tables, where both the causes $(X_1, \ldots, X_n)$ and the effect (E) are binary variables. Now, suppose you know you are in a fully observed and deterministic setting and have observed outcomes that cannot be explained by any small set of variables. Determinism demands that outcomes be perfectly explainable without recourse to randomness or noise, so a rational observer will invoke as many variables as necessary (i.e. all variables except for the outcome variable in Figure 1 left panel). However, if you have not observed all the variables in the setting, it could still be that a small set of the variables you have observed are causative of the outcome, while some of the hidden variables may also be the effective causes.

To further illustrate, suppose a learner has a dozen encounters with a system involving three binary inputs X_1 to X_3. Suppose every outcome but one can be explained by the presence or absence of X_1; but the one remaining "outlier" can then only be explained via positing a complicated conjunction of all three observable variables. With no hidden variables, the only rational solution would be to posit this complex explanation. However, if we allow for even one hidden variable H_1 to exist, it becomes possible that the outlier could simply result from the action of the hidden variable, leading to a simple explanation in terms of the observable features. That is, a rational explanation may evoke only X_1 explicitly and implicitly marginalizes over a hidden complicating possibility, the unobserved states and involvement of H_1. As the number of hidden variables increases it becomes increasingly plausible that the outcome could be entirely due to some combination of hidden variables, increasing the credence that a normative learner should give a "null hypothesis" type explanation where none of the measured variables matter at all for the outcome: at least implicitly the randomness of the outcome is driven by the increasing chance of it being controlled entirely by unobserved causes.

X_1	X_2	E
0	0	0
0	1	0
1	0	0
1	1	1

AND:
0001, 1000
OR:
0111, 1110
XOR:
0110, 1001
Singular:
0011, 1100, 0101, 1010
One Exception:
0010, 1011, 0100, 1101

Figure 1: Stimuli tested in the paper. Each trial involved four observations that show the outcomes of four combinations of X_1 and X_2. Fourteen stimuli were classed into five categories given their logical class. `Singular` stimuli can be explained by a single observed variable; the `One exception` group includes stimuli that cannot be explained in terms of the observable variables using any of the standard two place Boolean combinators. The brown color indicates stimuli that must invoke negation to explain outcomes under the `AND`, `OR`, and `Singular` categories.

We will first introduce our normative model, and use simulations to demonstrate the idea above. We will then conduct an empirical experiment to see whether people exhibit the a similar kind of sensitivity to latent complexity as the rational account.

A program induction approach to express causal rules

In this section, we demonstrate how explanations in terms of observable variables should shift from complex to simple as the number of hidden variables in a situation increases. We adopt a Bayesian approach where the hypotheses and prior are defined via program induction. We work with a deterministic likelihood such that a complete mechanistic explanation can only be correct if it fully explains all observations.

Probabilistic context-free grammars

We use `AND` ($\wedge$), `OR` ($\vee$), and `NOT` ($\neg$) as basic primitives to express causal rules. For example, $(X_1 \wedge X_2) \vee (\neg X_3)$ means that for the effect E to appear, you need either the combined presence of X_1 and X_2, or the absence of X_3. To sample explanations, we use a simple disjunctive normal form grammar (DNF) as outlined in Goodman et al. (2008) and Buchanan et al. (2010) [1]:

$$S \rightarrow \forall x,\ l(x) \Leftrightarrow (D) \tag{1}$$

[1] A choosing of primitives and grammar forms means that our model is prescriptive respecting the higher-level framework, but its specific predictions will vary given specific primitives and grammars.

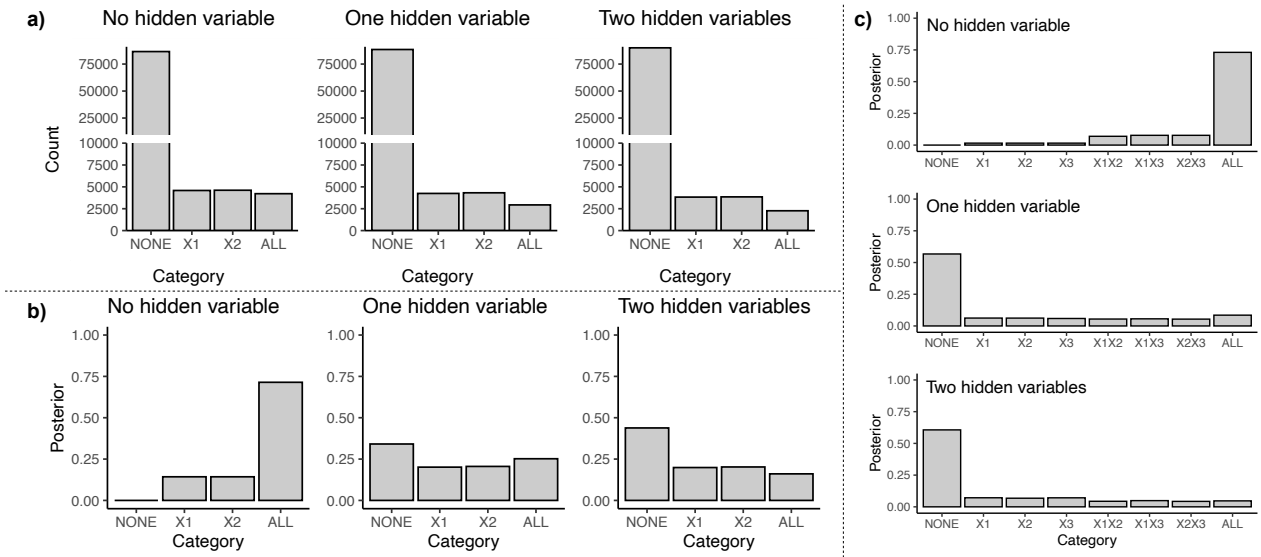

Figure 2: Model simulation results. a) The distribution of simulated hypotheses in each category under a sample of 100,000. b) Posterior distributions averaged from stimuli in Figure 1. c) Posterior distributions averaged from all possible stimuli for three explicit variables.

$$D \rightarrow (C) \vee D \tag{2}$$

$$D \rightarrow False \tag{3}$$

$$C \rightarrow N \wedge C \tag{4}$$

$$C \rightarrow True \tag{5}$$

$$N \rightarrow \neg N \tag{6}$$

$$N \rightarrow P \tag{7}$$

$$P \rightarrow X_1 \ldots X_n \tag{8}$$

For each sampled hypothesis, the generative mechanism initializes a disjunctive form placeholder (D) and iteratively replaces the terms according to the production rules above until all terms in the expression reach their terminal status ($True$, $False$, or one of the variables from X_1 to X_n). We apply the principle of indifference, assuming that during each replacement process, each valid production rule has equal probability of application. Production rules (6) and (7) result in a higher probability for a generative relationship than a preventative relationship (i.e. direct assertions are the default and an additional negation step is required to invert them making negative rules less likely to be generated than positive). Of note, we chose to do this rather than treating positive and negative assertions as symmetric because this both reflects the intuition that negation increases an assertion's complexity and aligns with empirical findings that generative relationships are easier to discover than preventative ones especially when the effect is assumed to be absent by default (Gong & Bramley, 2023; Cheng, 1997). Future work will investigate the extent to which this affects predictions.

Likelihood and the hidden variables

The approach outlined above allows us to generate a prior sample of hypotheses that are naturally biased toward simplicity, and in the limit of infinite sampling would cover all expressions in propositional logic relating the variables to the outcome. Since we assume a deterministic setting we can simply "filter" on this sample, ruling out all models that cannot explain the observation to arrive at a posterior sample and marginalize over this to calculate posterior probabilities for the involvement of different variables.

When there are hidden variables, those variables whose presence or absence is unknown are included in $X_1, \ldots, X_n$ as well, and the likelihood is calculated by marginalizing over all possible combinations of states of the hidden variables (assuming uniform priors on whether they are present or absent on each trial). For example, if there are two hidden variables, the likelihood would marginalized (averaged) over the four states of presence and absence of the two hidden variables ({0,0}, {0,1}, {1,0}, {1,1}).

Clustering the hypotheses

In the sampling process, we will find that many hypotheses are syntactically different but semantically identical (e.g., both $X_1 \wedge X_2$ and $X_2 \wedge X_1$ express a conjunction between X_1 and X_2 at the semantic level). We also note that rules can be harder to articulate, especially when the rule is complicated or involves hidden variables. Therefore, instead of focusing solely on the syntax or semantics, we concentrate on a higher level by clustering hypotheses into different categories depending on which observed variables they involve.

When there are two explicit variables, X_1 and X_2, we identify four categories: Rules in which (1) Neither of X_1 and X_2 are relevant; (2) only X_1 is relevant; (3) only X_2 is relevant; (4) both X_1 and X_2 are relevant. These categories are referred to as "NONE", "X1", "X2", and "ALL" throughout the rest of the paper. Each hypothesis belongs to one of these categories based on whether the outcome predictions can depend on the state of X_1 or X_2.

Simulation

Figure 2a illustrates the distribution of hypotheses across different categories when sampling 100,000 hypotheses, which will serve as the priors. The majority of hypotheses fall into the "NONE" category, with the fewest falling into the "ALL" category: expressions involving only TRUE or FALSE have higher probabilities to be generate than those involving $X1$ or $X2$. This distribution aligns with the intuition that explanation complexity increases in the order: NONE<X1=X2<ALL; the "ALL" category involves longer expressions that encompass relationships invoking roles for both X_1 and X_2.

A dominance in prior does not imply that "NONE" is always the best answer. With no hidden variables, the "NONE" category can only explain observations where the effect always appears or never occurs, irrespective of X_1 and X_2. We tested the model's predictions on 14 types of observations where the outcome does not remain constant (see Figure 1), a set that will be empirically validated later. As shown in Figure 2b, when no hidden variables are present, explanations often require both X_1 and X_2 to account for the observations. However, this dynamic shifts as the number of hidden variables increases. The posterior probability of the "NONE" category increases, along with those of the "X1" and "X2" categories. It implies that complex situations can more plausibly be attributed fully or partially to the influence of hidden variables. Consequently, our best explanations, regarding the role of the observed variables, revert to simpler ones.

Meanwhile, the role of hidden variables could be significant or minimal depending on the types of stimuli. If the ground truth of a stimulus is already simple (i.e., the prior is already high), the pressure to leverage hidden variables is low. Figure 5 shows the priors of five different types of stimuli (in the strip) and the model predictions (marked as X). The differences among no, one, and two hidden conditions increases when the ground truth prior decreases (from top to the bottom). This means the model only makes the switch when the explanation would otherwise be highly complex.

Figure 2c shows how the model predictions extend to scenarios involving three explicit variables. Although in this simulation section we applied a limited number of primitives and variables, future work will test it on more complex situations that involve different primitives and more variables.

Experiment

Methods

Participants 90 participants (31 female, 58 male, 1 preferred not to say, aged 41±11) were recruited via Prolific Academic and were paid £2. The task took around 18 minutes. The anonymous data and analysis code, as well as experiment procedure are available (https://osf.io/5yarw/).

Design Participants were asked to imagine themselves as "medical alchemists" who needed to determine the roles of different ingredients in producing medicine (Figure 3a). For each trial, they observed the brewing process of four potions,

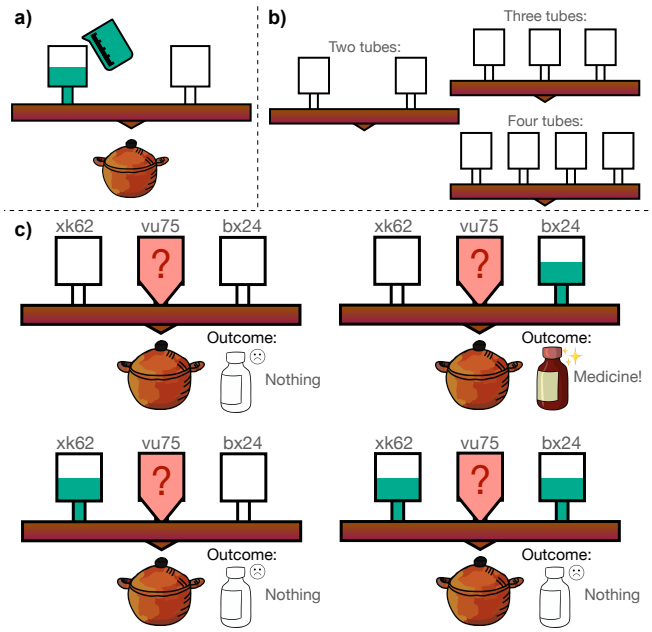

Figure 3: The cover story and stimulus example. a) The potion and equipment. b) Different number of tubes. c) One experimental trial that represents "0100" situation in Figure 1.

which were identical before any ingredients were added. Ingredients were added through equipment that could contain two, three, or four tubes (Figure 3b), and then the outcome showed whether the potion successfully became medicinal or not (Figure 3c). Participants were further instructed that some tubes might be covered with a red cloth (Figure 3c), indicating that participants would not know whether corresponding ingredients were added to each potion or not.

As a first foray, we only tested the setting with two explicit variables and 14 stimuli (Figure 1). We examined three conditions where the number of hidden variables varied between 0, 1 and 2. This implies that equipment with two, three, or four tubes would always have zero, one, or two tubes covered by red cloth, respectively. The conditions and stimuli were arranged using the Latin square design so that each participant experienced all observed patterns, but only in one of the three hidden complexity conditions. This resulted in three stimulus lists to which participants were randomly assigned.

After viewing the four observations for each trial, participants were asked a forced-choice question: Which one of the following statements best reflects the truth about the two focal causes: Neither of X_1 and X_2 is relevant; only X_1 is relevant; only X_2 is relevant; both X_1 and X_2 are relevant. Here, X_1 and X_2 correspond to ingredient names that varied from trial to trial. Participants were also asked to provide a confidence rating for their response on a 0-100 scale.

Procedure Before beginning the task, participants were instructed about the cover story, the meaning, and the examples of "relevance" (generation, prevention, or rules that combine it with other ingredients). The deterministic setting was emphasized by explaining that if the same ingredients are added through the equipment, the result will always be the same.

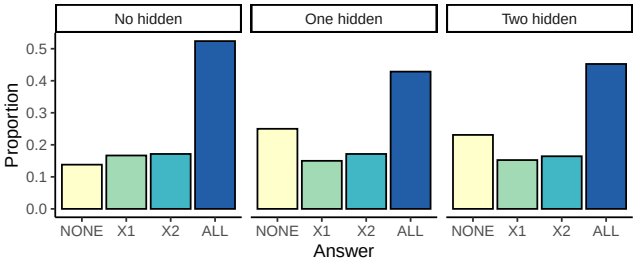

Figure 4: The proportion of participants' answer for each option.

Participants had to pass comprehension check questions before the task. The order of the four observations in each trial and the positions of tubes covered by red cloths were randomized among trials. The order of trials and the order of four forced choices were randomized among participants.

Results

Figure 4 shows participants' answer for each category under three conditions with different numbers of hidden variables. Three answer distribution differed (χ^2 test of independence: $\chi^2(6) = 19.52$, $p = .003$). The distribution of answers significantly differed between "No Hidden" vs. "One Hidden" ($\chi^2(3) = 17.92$, $p < .001$), or "No Hidden" vs. "Two Hidden" ($\chi^2(3) = 12.34$, $p = .006$), but not "One Hidden" vs. "Two Hidden" conditions ($\chi^2(3) = 0.66$, $p = .88$). When hidden variables were present, the proportion choosing "NONE" increased, while the proportion choosing "ALL" decreased, which reflects the previous simulation results (Figure 2b).

Proportions by stimulus types Figure 5 shows participants' answers broken down by type of stimuli. For each type of stimuli, we focus on whether participants chose none, all, or one of the variables as related (i.e. "X1" and "X2" are merged as the category "ONE" here). The stimulus types are ordered according to how likely the corresponding deterministic observable variable explanation (Figure 1) would be sampled from the PCFGs. This means that any prior sampled hypothesis would be more likely to fall under `Singular` ground truth than `AND` ground truth, and so on. For each type, the answer distributions only significantly differed between conditions for `OR` ($\chi^2(4) = 13.37$, $p = 0.01$, Fisher's exact test was used to due to small numbers in some cells: $p = 0.006$) and `One exception` ($\chi^2(4) = 13.08$, $p = 0.01$) stimuli. These two types also have lower priors than all other types except for the `XOR`. This is aligned with our model prediction that when the situation is more complex, people will turn to simpler explanations whenever it is possible. [2] We will later discuss why we think `XOR` was an exception.

The one-exception cases We finally explore the patterns for `One exception` stimuli. They have either only one presence ($\{0,0,1,0\}$, $\{0,1,0,0\}$) or only one absence in the outcome ($\{1,1,0,1\}$, $\{1,0,1,1\}$). Preferences seemed to differ between this two types. When facing one-present stimuli, participants, especially in the one-hidden condition, tend to refer to one variable rather than select "ALL" or "NONE" (Figure 6). This reflects that in the situation when a singular variable can explain all but one data point (e.g. X_1 can fully explain $\{0,0,1,0\}$ as long as we turn the fourth outcome from 0 to 1), participants may leverage the hidden variable in the system to help maintain a singular and generative answer, an answer that is more informative than "NONE" and less complex than "ALL". This tendency was not statistically significant, partially because of the sample size. This needs to be further tested in future work, also including situations that involve more explicit variables.

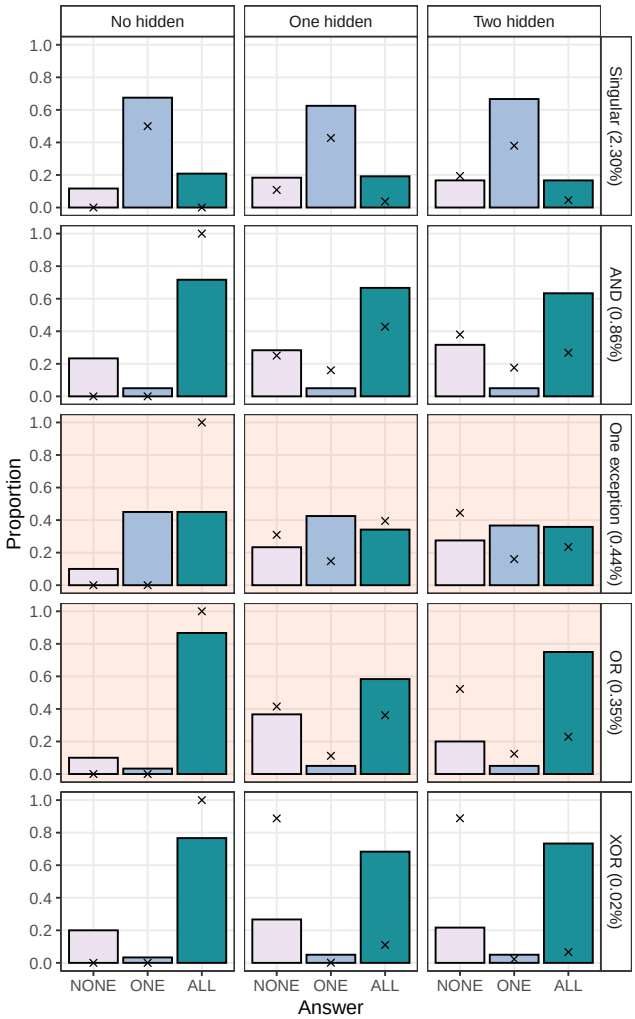

Figure 5: Participants' answer for under each stimulus type. Model predictions are marked as "X". Numbers in the strips indicate the stimulus probabilities in the prior. Coral color indicates distributions significantly differ between conditions.

Confidence Participants confidence ratings differed between conditions ($F(2,89) = 12.04$, $p < .001$). This was significant between the no-hidden condition (75 ± 18) and one-hidden condition (68 ± 19, $t(89) = 4.12$, $p < .001$), or the no-hidden condition and two-hidden condition (66 ± 21, $t(89) =$

[2] It is worth noting that because we did not add parameters in our model to help fit with human data, we focused more on qualitative patterns. Nonetheless, we noticed quantitative deviations between humans and models for `OR` and `One exception`, which are open to future investigation.

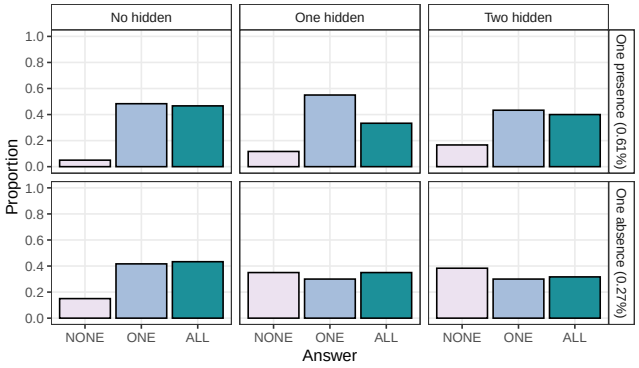

Figure 6: Participants' answers in two subsets of stimuli under the "one expection" category.

4.84, $p < .001$), but not between two hidden variable conditions ($t(89) = 1.42$, $p = .48$). This is consistent with previous results where the answer proportions did not significantly differ between one hidden variable and two hidden variables (Figure 4). Participants' confidence ranked from stimuli `AND` (75 ± 24), `Singular` (70 ± 23), `OR` (70 ± 25), `XOR` (68 ± 25), to `One exception` (67 ± 26, $F(4,89) = 3.90$, $p = .006$). Only the difference between `AND` and `One exception` was statistically significant ($t(89) = 3.80$, Bonferroni-adjusted $p = .003$).

Discussion

In this paper, we provide a model that demonstrates how type-level explanations should differ their complexity in environments involving varying numbers of hidden variables. By leveraging a program induction approach, our model allowed for consideration of a wide variety of belief hypotheses, reflecting the reality that the causal explanations people could form in principle is practically unbounded. With this approach, we demonstrate that complex beliefs are demanded when the environment is fully observed yet lacks a simple explanatory mechanism, while simpler explanations persist when the environment is more uncertain.

Our model provides a new insight into a long-standing philosophical question: How to choose among theories. It also speaks to the modern statistical question of how scientists should choose between models (Doroudi & Rastegar, 2023). A good theory or model should not only be simple but also informative in explaining phenomena (Jefferys & Berger, 1992; Pacer & Lombrozo, 2017). We demonstrate how this tendency towards simplicity and informativeness is reflected in a Bayesian framework, via the prior and likelihood respectively, and how Bayes' rule can help guide the combination of the two to provide a balanced answer.

With the model simulations, we identified two normative principles that could be tested empirically: (1) When there are hidden variables, a learner should shift from complex explanations toward simpler explanations; (2) this shift should be more pronounced when the only fully observable explanation for the pattern is more complex (i.e., has a low prior). We showed both these principles are reflected in human judgment tendencies. Participants were more inclined to choose simpler explanations when there were hidden variables, and their response distributions were more varied across conditions for the `One exception` and `OR` stimuli in line with their lower priors.

We also observe deviations between human performance and our account. For example, the second principle mentioned above did not manifest for `XOR` stimuli which received the lowest prior under our PCFGs. The low prior is a consequence of the grammar not containing `XOR` as a primitive (having to construct it by combining `AND`, `OR` and `NOT`). Although `XOR` has historically been treated as complex and difficult case due to its non-monotonicity, particularly in the early connectionist literature, recent research suggests that `XOR`, representing "either A or B", is a salient possibility for that human reasoners will readily entertain (Jiang & Lucas, 2024; Bramley & Xu, 2023; Gerstenberg & Icard, 2020). Therefore, future work will attempt to adapt the primitives and architecture of the model to explore what better reflects human inductive biases.

The other two deviations are suggestive of how human interpretation of the task differs from the assumptions based into the normative model. First, in our experiment, when no hidden variable exists, a rational learner should never choose "NONE", given that the outcomes in our stimuli always change according to the changing variable states. However, the "NONE" category still received 14% of answers (Figure 4). Second, compared to the model, participants were more likely to choose "ALL" (Figure 4 vs. Figure 2b), and they were relatively insensitive to the difference between one and two hidden variables. The reasons behind these deviations could be multifaceted. It could be due to the task interface, where we did not emphasize that the masked ingredients were chosen randomly. This may affect how people consider the probability that the hidden variables would play a role in interpreting the outcomes, as well as the probability that the hidden variables would be present in each scenario. Besides, human cognition may deviate from an unbounded rational model due to resource considerations (Lieder & Griffiths, 2020). Instead of exhaustively incorporating all possible states of hidden variables in the inference process, people may use heuristics or more computationally economical approaches, e.g., ignoring the larger, unknown space of things unless they are necessary to explain the observation (Gershman, 2019). Future work could improve the experimental instructions, and examine how cognitive resource limitations factor into this process.

Finally, it is worth noting here that the model provided in this paper is not intended to be a universal rule applicable to all scenarios. In scientific research or everyday causal discoveries, there may be other considerations. For example, an explanation could often not just be a static conclusion but would serve as guidance for future information search. Future work could integrate the roles of future investigation goals in shaping the learners' current explanations.

Acknowledgement

TG, CL and NB were supported by an EPSRC New Investigator Grant (EP/T033967/1), United Kingdom. SV was supported by a Principal's Career Development Scholarship, awarded by the University of Edinburgh and received funding from the Institute for Language, Cognition and Computation at the University of Edinburgh.

References

Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing neurath's ship: Approximate algorithms for online causal learning. *Psychological Review*, *124*(3), 301–338.

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.

Bramley, N. R., & Xu, F. (2023). Active inductive inference in children and adults: A constructivist perspective. *Cognition*, *238*, 105471.

Bramley, N. R., Zhao, B., Quillien, T., & Lucas, C. G. (2023). Local search and the evolution of world models. *Topics in Cognitive Science*.

Buchanan, D., Tenenbaum, J., & Sobel, D. (2010). Edge replacement and nonindependence in causation. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 32).

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367–405.

Doroudi, S., & Rastegar, S. A. (2023). The bias–variance tradeoff in cognitive science. *Cognitive Science*, *47*(1), e13241.

Fränken, J.-P., Theodoropoulos, N. C., & Bramley, N. R. (2022). Algorithms of adaptation in inductive inference. *Cognitive Psychology*, *137*, 101506.

Gershman, S. J. (2019). How to never be wrong. *Psychonomic Bulletin & Review*, *26*, 13–28.

Gerstenberg, T., & Icard, T. (2020). Expectations affect physical causation judgments. *Journal of Experimental Psychology: General*, *149*(3), 599–607.

Gong, T., & Bramley, N. R. (2023). Continuous time causal structure induction with prevention and generation. *Cognition*, *240*, 105530.

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

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Griffiths, T. L., & Tenenbaum, J. B. (2007). Two proposals for causal grammars. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 323–345). New York: Oxford University Press.

Jefferys, W. H., & Berger, J. O. (1992). Ockham's razor and bayesian analysis. *American Scientist*, *80*(1), 64–72.

Jiang, C., & Lucas, C. G. (2024). Actively learning to learn causal relationships. *Computational Brain & Behavior*, 1–26.

Johnson, S. G., & Ahn, W.-k. (2015). Causal networks or causal islands? the representation of mechanisms and the transitivity of causal judgment. *Cognitive Science*, *39*(7), 1468–1503.

Johnson, S. G., Valenti, J., & Keil, F. C. (2019). Simplicity and complexity preferences in causal explanation: An opponent heuristic account. *Cognitive Psychology*, *113*, 101222.

Kushnir, T., Gopnik, A., Lucas, C. G., & Schulz, L. (2010). Inferring hidden causal structure. *Cognitive Science*, *34*(1), 148–160.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, 1–60.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, *55*(3), 232–257.

Lombrozo, T. (2010). Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*(4), 303–332.

Lucas, C. G., Holstein, K., & Kemp, C. (2014). Discovering hidden causes using statistical evidence. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual conference of the cognitive science society* (pp. 892–897).

Luhmann, C. C., & Ahn, W.-k. (2007). Buckle: a model of unobserved cause learning. *Psychological Review*, *114*(3), 657–677.

Pacer, M., & Lombrozo, T. (2017). Ockham's razor cuts to the root: Simplicity in causal explanation. *Journal of Experimental Psychology: General*, *146*(12), 1761–1780.

Piantadosi, S. T. (2021). The computational origin of representation. *Minds and machines*, *31*, 1–58.

Rottman, B., Ahn, W.-k., & Luhmann, C. (2011). When and how do people reason about unobserved causes. In P. Illari, F. Russo, & J. Williamson (Eds.), *Causality in the sciences* (p. 150–183). OUP Oxford.

Sulik, J., van Paridon, J., & Lupyan, G. (2023). Explanations in the wild. *Cognition*, *237*, 105464.

Valentin, S., Bramley, N. R., & Lucas, C. G. (2022). Discovering common hidden causes in sequences of events. *Computational Brain & Behavior*, 1–23.

Zemla, J. C., Sloman, S. A., Bechlivanidis, C., & Lagnado, D. A. (2023). Not so simple! causal mechanisms increase preference for complex explanations. *Cognition*, *239*, 105551.

Zhao, B., Lucas, C. G., & Bramley, N. R. (2024). A model of conceptual bootstrapping in human cognition. *Nature Human Behaviour*, *8*, 125–136.