CHAPTER 3

Quantifying explanation preference in a probabilistic context

3.1 Abstract

People have an intuitive feeling about on how satisfying an explanation is. In past work, a "simplicity preference" for certain explanations has been argued to be a major consideration in how people prefer some explanations over others. I designed a new experimental paradigm that more clearly shows how the prior (probability of the cause being existent) and causal strength (probability of effect happening given the cause being existent) of a causal system can affect people's overall preference for simple or complex explanations. I found that instead of being a universal preference, a simplicity preference for explanation is only present when either the prior or the causal strength of this factor is distinctively high. Moreover, a standard Bayesian posterior estimation of the posterior of some explanation being true is less descriptive of the empirical data compared to a heuristic model which indicates that people feel most satisfied with the explanation including all the causes with distinctively high prior and causal strength causes, but not more redundant than that.

3.2 Introduction

People are able to make a judgment about how good an explanation is. In daily life people intuitively feel whether an explanation "makes sense" (Is this why my kid got sick? Is that a sufficient reason for paying so much for this service?). On a cultural level, people unsatisfied with mainstream explanation regarding certain political issues may resort to conspiracy theories, which appears to be a more compelling explanation. This peculiar feeling of "satisfaction for explanation" is very common and critical on every level of life, thus it is important to better characterize and understand it.

In search of the origin and the characteristics of the feeling of explanatory satisfaction, one

helpful starting point is the philosophical study of scientific explanation. A scientific theory regarding certain empirical phenomenon is basically an explanation for that observation which the scientific community collectively agree on. Throughout the history of science there have been plenty of times when a once orthodoxical theory (e.g., Newtonian mechanics, phlogiston) is replaced by better ones (e.g., quantum mechanics, oxygen theory of combustion). Thus, every scientist needs to be able answer these questions: what makes one explanation more satisfying the other? How can I be confident that my theories and models are better than the existing ones?

For example, chemist Lavoisier explained his criticism against the phlogiston theory as an explanation of combustion phenomenon as:

> If all of chemistry can be explained in a satisfactory manner without the help of phlogiston, that is enough to render it infinitely likely that the principle does not exist, that it is a hypothetical substance, a gratuitous supposition. It is, after all, *a principle of logic not to multiply entities unnecessarily.*

These statements are closely related to the notion of Occam's razor, stating that other things being equal, simpler theories are better (for more philosophical discussion on simplicity and quotes from the history, see Baker, 2016). Besides simplicity, researchers have proposed other features that make an explanation preferable, i.e., explanatory virtues, such as complexity (Zemla *et al.* 2017; Lim & Oppenheimer 2020, especially for more complex phenomena), coherence (Thagard 1989), unification (Myrvold, 2003), and so on. Many of these virtues are not only concepts proposed by philosophers, but also have been empirically shown to be factors that people actually use to justify their preference for explanations (Zemla *et al.*, 2017). But are these explanatory virtues the most logical explanation of explanation preference? Could there be more fundamental principles underneath these individual factors?

Among all the explanation virtues, the factor of simplicity and complexity has received much attention from researchers in philosophy, psychology and cognitive science. As Lavoisier's quote points out, a good explanation should only include necessary causes but no more. If this is also a general cognitive preference for people beyond scientist, it may be stated as: people should prefer an explanation that is complex enough to be able to account for the explanandum, yet simple enough to be more probable and not redundant.

However, it is not obvious how to specify this trade-off in a quantitative and behaviorally testable manner.

There are at least two paths to deal with this problem. One way is to study what are the conditions make people prefer either a simple or complex explanation (defined in terms of the causal structure, most often the number of causes, but see an exception in Pacer & Lombrozo 2017). Experiments have been designed to manipulate conditions such as probability (Lombrozo, 2007),

stochasticity (Johnson *et al.*, 2019), existence of mechanism (Zemla *et al.*, 2020) and knowledge domain of the causal story (Johnson *et al.*, 2019) to probe when those preference shift towards more simplicity or complexity. However, these kinds of methods usually only address the simplicity preference and the complexity preferences separately. Looking into the related manipulations is helpful in exploring different motivations for either simplicity and complexity preference, but this method could not answer: is there a single mechanism or algorithm that could unify these phenomenon, for example, in terms of maximizing some computational quantity?

Alternatively, a probability modeling perspective can be adopted as a framework to organize different factors. This has been more often seen in psychology studies of causal inference (Griffiths & Tenenbaum, 2009; Lu *et al.*, 2008), where the causal structure is usually treated as a probabilistic Bayesian network with prior probabilities for each independent causal node and dependent probability distributions to demonstrate the relations between causal effects (see an example network in Figure 3.1 which we will further explain in Section 3.3). Formal models for evaluating the quality of explanation have been developed, such as the most probable explanation (probable in terms of posterior, see Pearl 1988) and the most relevant explanation (relevant in terms of likelihood, see Yuan *et al.* 2011). People's judgement of preferred explanation can then be contrasted to the model prediction, determining which model best describes human preference (Pacer *et al.*, 2013). This path of research is not only useful in quantifying human explanation preference but also can be insightful for Artificial Intelligence (AI) systems to be explainable (Madumal *et al.*, 2020). However, it is unclear how to connect those formal models with the literature on explanatory virtue since the models are not described in those terms.

In fact, conceptually, the Bayesian probabilistic modeling could be interpreted in terms of balancing the simplicity and complexity of a model or explanation. The benefit of complexity to cover empirical observations can be interpreted as likelihood, i.e., $P(\text{Observation}|e)$, where higher likelihood means better chance of generating such observation. The simpler explanations, on the other hand, are a priori more plausible, i.e., having higher prior $P(e)$. The balance between complexity and simplicity could then be easily put as maximizing the posterior which is proportional to likelihood multiplied by the prior of a given explanation.

Previous empirical research has provided some indirect support of this view. Lombrozo (2007) found that people are biased towards simpler explanations because they over-estimate the prior probability of simple explanation being present. If people are aware that the simple and complex explanations both have equal prior and equal likelihood (in this study, both are deterministic), then people no longer have a simplicity preference. Regarding likelihood, Johnson *et al.* (2019) found that when causes have lower likelihood, people tend to adopt more complex explanations so that the likelihoods adds up to be higher (we will explain this mathematical intuition in the Method section in Experiment 1). But none of these studies are actually formulated in these posterior probability

terms. Neither do the experimental materials provide sufficient information to test such theories. The one exception is Pacer & Lombrozo (2017) where they compared formal models and found that, in fact, the theories resorting to posterior maximization do not explain human data well. Note that this was only tested on two specific causal structures with specific probabilistic parameters.

Thus, for the current study, we have two goals: first, to design an experimental paradigm that allows more precise and explicit manipulation regarding prior and likelihoods of the causal system. This will allow us to test the simplicity preference with a larger range of possible stimuli than most of the previous studies. Second, rather than simply looking at the ratio between choosing simple versus complex explanation, we plan to apply quantitative models to compare with behavioral data, further exploring whether there could be a theory framework to account for the explanation preference under different conditions.

3.3 Research goal

Our goal is to develop a novel behavioral paradigm which enables us to characterize people's explanation preference with computational models.

On the theory side, we will start by testing theories including:

- simplicity preference: people have a bias to select simpler explanations.
- complexity preference: people have a bias to select more complex explanation when the causes stochastically lead to the effect (instead of deterministic).
- posterior preference: people judge the quality of explanation by its posterior probability, which is a combination of prior and causal strength.

On the experimental design side, to test the theories above, we have to allow the causal structure to be probabilistic (stochastic), therefore we choose to use Bayesian network to design the context of explanation. To allow people to choose between explanations that vary in their simplicity, we specifically choose the common effect causal structure (also called "collider structure"), where causes are independent from each other and each could contribute to the existence of the effect (Figure 3.1). In this way, simplicity of an explanation can be defined as the total number of causes since there is no hierarchical structure between the causes (unlike Pacer & Lombrozo 2017). Note that this is a relatively arbitrary choice for our initial exploration: other alternatives such as common cause structure or hierarchical structures can easily be adopted into our paradigm.

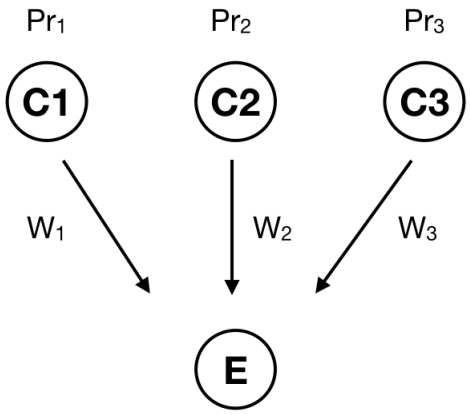

Figure 3.1: We use the common effect structure (also called collider structure) for our experiments. Causes C1, C2 C3 are independent from each other, with a prior probability Pr_1, Pr_2. Pr_3. Each cause probabilistically cause the effect E, with the probability $P(E|C_i, \neg C_{j\neq i})$ denoted as w_i.

3.4 Overview of experimental procedures

All of the experiments are under the cover story of "deciding the best explanation for an alien patient's symptom". This cover story has been used in the previous literature on explanation preference (see Lombrozo 2007).

At the beginning of the experiments, participants watch a video with an alien narrator "Doctor Luzeka" from "planet Omega" introducing the task. Our key aim of this instruction is to explain that causes are independent from each other and with some probability causing the same effect. We used the familiar analogy "hot pepper, colds and allergies independently cause sneezing". Furthermore, it is emphasized that "how common a cause occurs could differ" (i.e., prior) with the example that allergy might be more rare than cutting chilli pepper given the cultural preference. Also, we highlight that "causes with different strengths are independent but could be added up" with the example saying that:

It is not the case that both of cold or allergies have to be present for sneezing to happen. Rather, having any one of the causes can already make Emily sneeze. But of course, if Emily both catches a cold and has seasonal allergies, then the chance of sneezing will increase.

The full transcript of the tutorial narration is shown in Appendix 3.8.1.1.

After the tutorial, there are three sections of three different types of questions. In the first two sections, participants are asked to provide some judgements regarding *how often some medical conditions occur* and *the effects of the combination of several medical conditions*. These are for the participants to familiarize with reading the information about each cause as well as thinking about conjugations of multiple causes. For example, regarding the prevalence of the condition, that is, prior probability, the trial goes like:

Betisia, Ryi Disorder and Qetrophy are all diseases that appear independent of each other.

Then the description of each cause is being presented, such as:

Betisia is very common among the population: among 100 aliens, 80 of them may have Betisia.

Participants are asked to answer:

Among 100 aliens in random population, how many of them may have Betisia, Ryi Disorder ***but NOT*** *Qetrophy at the same time?*

This procedure is the same for the second section asking about causal strength. An example description of a cause is like:

Betisia is a weak cause of fast puchim: among 100 aliens who have Betisia, 30 of them may suffer fast puchim.

The question follows:

Among 100 aliens who have Bestisia, Ryi Disorder and Qetrophy at the same time, how many of them may suffer fast puchim?

The exact interface of these trials are shown in Appendix 3.8.1.2.

Then the third section presents the individual alien patient with symptoms, i.e., the explanation section. In each trial, we provided three potential causes for the symptom, the cause being either having certain diseases or being exposed to certain chemicals. An example narrative goes as follows:

Many factors contribute to the symptom of nisis bleeding: having diseases such as Rozi Syndrome, Hypiria and Zodophy all lead to nisis bleeding, independently.

Then the information from a "public health report" is shown for each cause. In Figure 3.2 we show an example report card. Note that the card is initially blank and requires the participant to click on it to "flip the card" and show all the relevant causal information. This design is to add more interaction to encourage active learning, as well as a method to check attention based on the clicking behaviors.

Below this information is the final judgment question:

Now we have an alien patient Aludu suffering nisis bleeding, which of the following explanation best explains Aludu's symptom?

Then three options are provided. In most trials except the attention check ones, the three options include explanations ranging from simple to complex, such as:

A-Aludu has only Rozi Syndrome

B-Aludu has only Rozi Syndrome and Hypiria

C-Aludu has Rozi Syndrome, Hypiria and Zodophy

Note that the three options are always in this set order, of "cause A", "cause A and B" and "cause A, B and C". Thus "cause 1" is always included in the three options. In the following text we will refer "cause A" as 1-cause, "cause A and B" as 2-cause, and "cause A, B and C" as 3-cause.

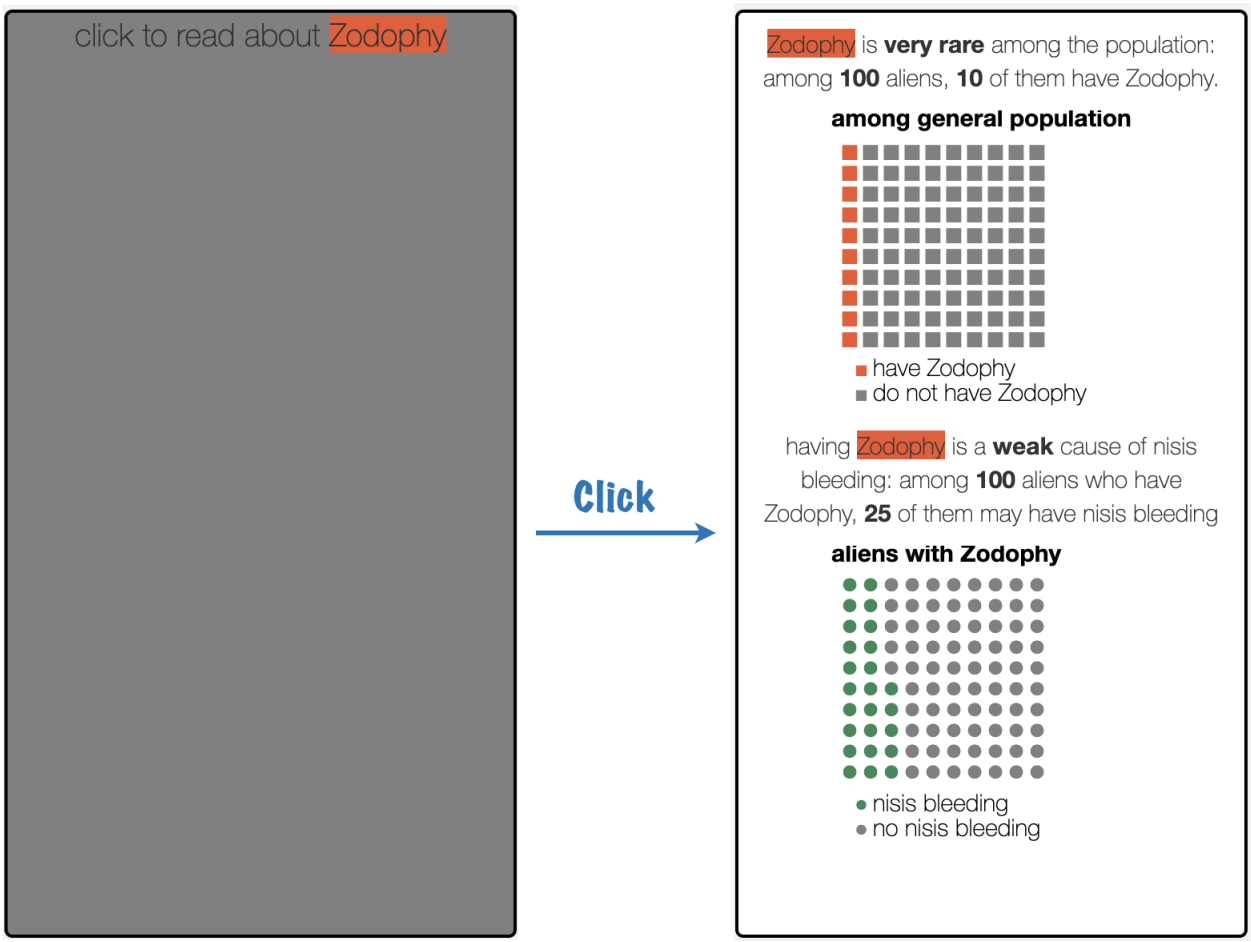

Figure 3.2: Example of the information card provided for a given cause. Initially the blank card (left) is shown on the screen. After participants click on the card, it flips and show the information card (right). On the information card, both the prior and causal strength information are presented in terms of both number and waffle graph.

After one explanation is chosen, an additional question appears on the screen saying *If you think the best diagnosis is not included above, please select the combination of causes that you prefer. Otherwise, please leave it blank and continue*. We added this free-choice option to allow participants directly expressing what they think is the best explanation.

3.5 Experiment 1

Experiment 1 is a preliminary test of people's explanation preference under clearly stated causal conditions. We contrasted a Bayesian posterior model of explanation evaluation with the behavioral data. We found this model has obvious deviation from the empirical data and proposed an alternative model.

3.5.1 Method

3.5.1.1 Bayesian posterior model for explanation

Because of the vast space of possible stimuli, we guided our stimuli design with a Bayesian posterior model, which has been suggested in previous research (Griffiths & Tenenbaum, 2009; Lu *et al.*, 2008; Pacer *et al.*, 2013). Specifically, in the current context of question, a more preferable explanation is the one with higher posterior given the observed data.

According to Bayes rule, the posterior is in proportion to the multiplication of prior probability of the explanation $P(e)$ and its likelihood $P(\text{data}|e)$. Now we specify both of these terms for the common effect causal structure.

First, for an explanation e that combines multiple causes $c_i, i = 1, 2, ...n$ which are mutually independent, the prior should be:

$$P(e) = \prod_i P(c_i) \tag{3.1}$$

Second, each individual cause is defined with a causal strength parameter w which represents the probability of inducing the effect $w = P(\text{effect} = True|\text{cause} = True, \text{other cause} = False)$ and a probability of failing to induce $P(\text{effect} = False|\text{cause} = True, \text{other cause}) = 1 - w$. To calculate the conjunctive causal strength, we adopt the "Noisy-OR" formulation (Cheng, 1997). The intuition of Noisy-OR rule is that the effect will only be *non-existent* if all the independent causes fail to have an effect. Formally, the cumulative causal strength of multiple causes is defined as:

$$P(\text{data}|e) = 1 - \prod_i (1 - w_i)^{E_i} \tag{3.2}$$

where E_i denotes whether the cause i is existent (1 means exist, 0 means not) and w_i is the causal strength of cause i. Since $1 - w_i$ is always between 0 and 1, the cumulative causal strength will monotonically increase with the number of causes present,

With the prior and causal strength, we can now compute the posterior for each explanation (in our setting, different combination of single causes):

$$P(e_j|\text{data}) = \frac{P(e_j) \cdot P(\text{data}|e_j)}{P(\text{data})} \tag{3.3}$$

To choose among the explanations, the probability of choosing each explanation is in propor-

tion to the posterior:

$$
\begin{aligned}
P(\text{choose } e_j) &= \frac{P(e_j|\text{data})}{\sum_{k=1}^{n} P(e_k|\text{data})} = \frac{P(e_j) \cdot P(\text{data}|e_j)/P(\text{data})}{\sum_{k=1}^{n} P(e_k) \cdot P(\text{data}|e_k)/P(\text{data})} \\
&= \frac{P(e_j) \cdot P(\text{data}|e_j)}{\sum_{k=1}^{n} P(e_k) \cdot P(\text{data}|e_k)}
\end{aligned}
\tag{3.4}
$$

This is a categorical distribution predicting the probability of choosing each option. We can then compare the empirical choice ratio to this distribution to evaluate the model performance (see Results section 3.5.2.1).

Note that the full set of alternative explanation e_k depends on the specific task setting. For example, in a forced choice task, the alternatives are all the options given by the question. In the free response task, all the possible alternatives are enumerated and considered.

3.5.1.2 Stimuli design

Since the Bayesian model of explanation preference is dependent on the prior and causal strength of individual causes, we designed the stimuli to separately test these two features.

For the first group of trials, the causal strength is fixed to be 0.9 across all causes while the prior changes at the level of 0.1, 0.5 to 0.9. We chose these prior levels because the model predicts very distinct explanation preferences in terms of choice ratio for each option. As shown in Figure 3.3B, model predicts that, from low to medium to high prior conditions, the most preferred explanation changes from the simplest (only one cause), to neutral (either one, two or three causes are equally preferred), to the most complex explanation, respectively.

Similarly, to test the effect of causal strength, in another group of trials the causes all have the same prior but their causal strengths are unequal. Again, for those trials the model exhibits distinctively different explanation preference, with each trial preferring either one, two or three-cause explanations (see Figure 3.4B).

Last, we designed another group of trials where all the explanations are similarly preferred according to the theory, i.e., having similar posteriors (see Figure 3.5C). This was intended to test if people have some preference beyond the probabilistic judgment (for example, Lombrozo 2007 found that people have a bias thinking simpler explanations to have bigger prior probability than the empirically presented probability information).

3.5.1.3 Participants

We recruited 78 participants (average age 39.2, standard deviation 10.7; 53 reported as males and 25 as females) from Amazon Mechanical Turk via Psiturk (Gureckis *et al.*, 2016). To only include

participants who truly read the relevant information into our analysis, we used the information card clicking behavior as the first exclusion criterion, since the information of each cause can only reveal when its card is clicked (see experimental procedure reviewed in section 3.4). Specifically, a participant will be excluded if they has one trial in the first 2 sections or has more than 1/5 of the trials in the explanation sections where at least 2 cards are not clicked. For those included participants, only trials with at least 2 cards revealed are considered valid thus included, otherwise the judgment of that trial is more likely a random guess rather than information-based.

In addition, there were attention check trials which have very obvious correct answer because the options to be chosen from are just the three single causes, one of them has highest prior or highest causal strength or both. Participants that make any mistake for the three test questions are excluded. Thus in total, we included 37 participants to the final analysis. All responses from these participants are valid trials in terms of the clicking behavior.

3.5.2 Results

3.5.2.1 Bayesian posterior model fails to predict the explanation preference

We compared explanation preference from the empirical data with the Bayesian posterior model prediction. Empirical responses are aggregated across participants to generate a proportion for each potential explanation, which could then be contrasted with the model probability output. Figure 3.3 shows that for the trials with different prior values, although the theory predicts very different patterns of simplicity/complexity preference across different levels of prior probability, the empirical data exhibit little sensitivity to the prior manipulation. Instead, a complexity preference for the 3-cause explanation is shared across all three trials regardless of the causes' prior probability.

Figure 3.4 shows that for the trials with different causal strength, the theory predicts two of the three conditions with correct ordering, indicating the effects of causal strength is somehow captured by the model. On the other hand, even though a simplicity preference was predicted for one condition (red line in Figure 3.4), participants still preferred the more complex 2-cause explanation.

Last, Figure 3.5 shows that for the trials with similar level of posterior according to the Bayesian model, the empirical data, however, still show an preference for the complexity (3-cause explanation).

The results above indicate that, compared to people, the Bayesian posterior model of explanation is overly sensitive to the prior yet has some degree of predictive power when causal strength is varied. Overall, the Bayesian posterior model does a poor job predicting the empirical preference for explanation.

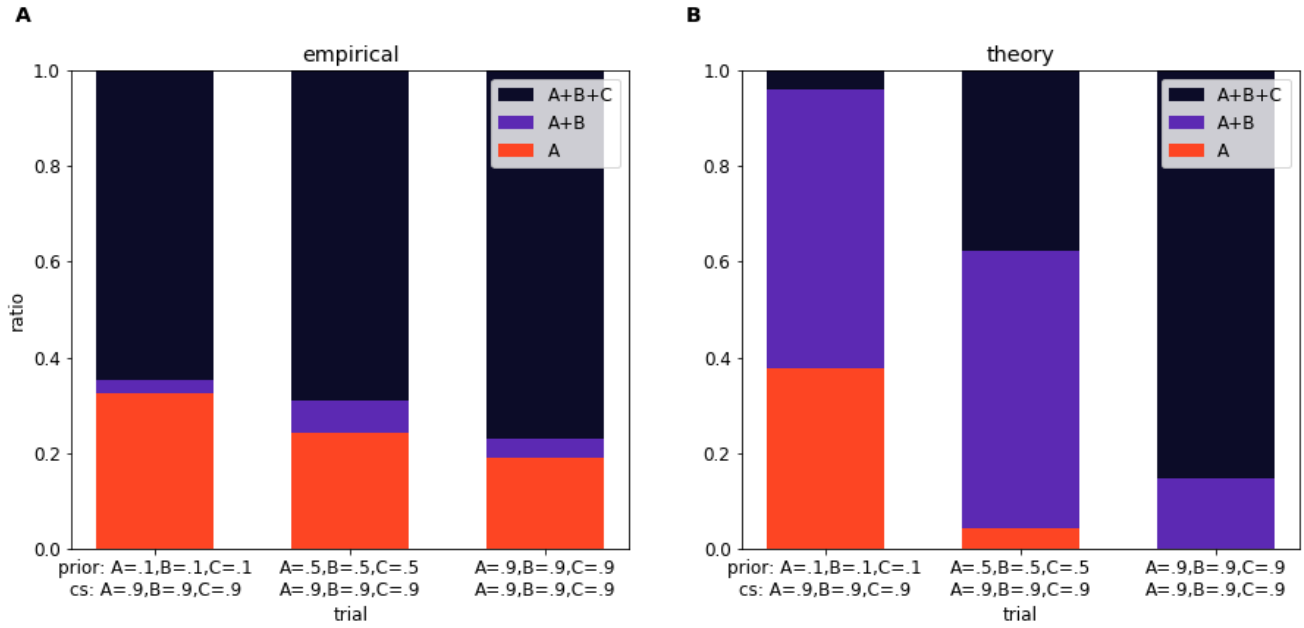

Figure 3.3: Choice ratio for trials with different prior and same causal strength. Left panel is the empirical average across all participants. Right panel is the theory predicted ratio based on the Bayesian posterior model.

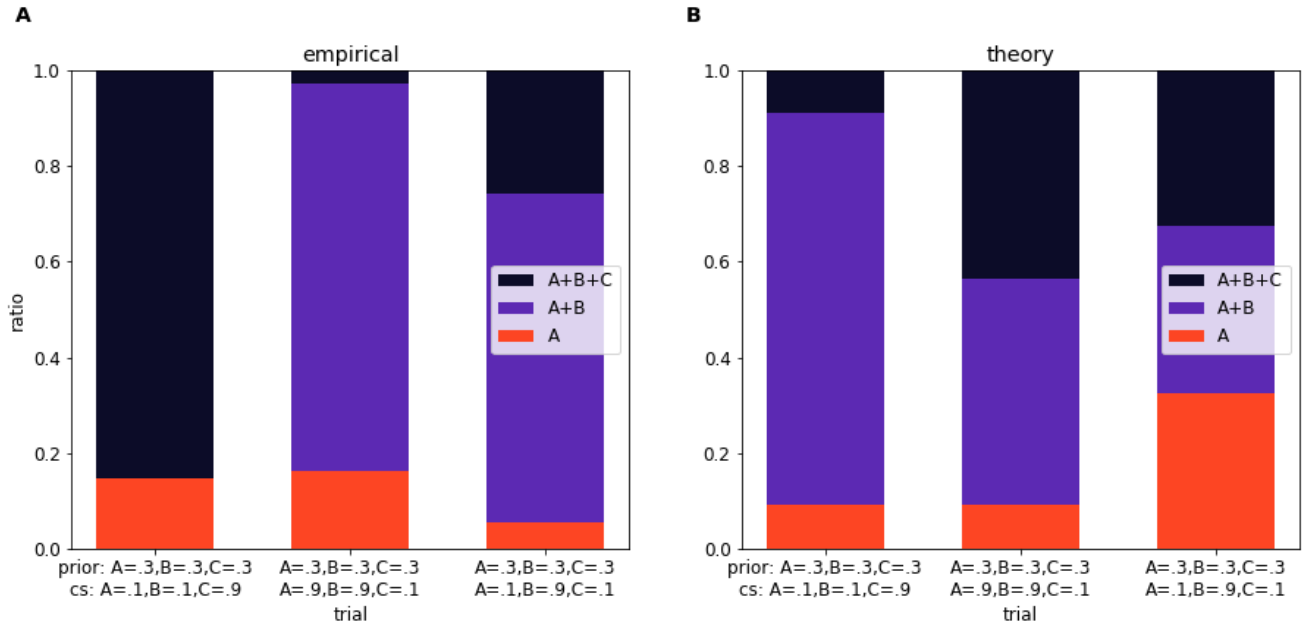

Figure 3.4: Choice ratio for trials with different causal strength and same prior. Left panel is the empirical average across all participants. Right panel is the theory predicted ratio based on the Bayesian posterior model.

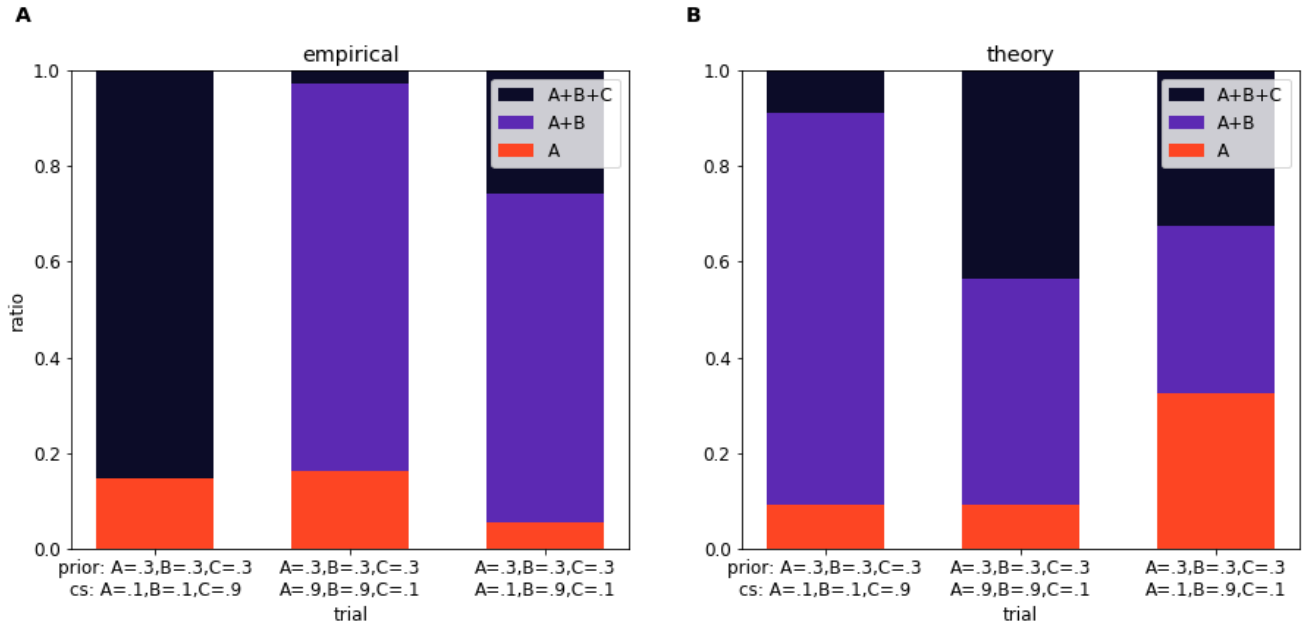

Figure 3.5: Choice ratio for trials with similar level of posterior according to the Bayesian model. Left panel is the empirical average across all participants. Right panel is the theory predicted ratio based on the Bayesian posterior model.

3.5.2.2 Heuristic model explains the explanation preference

After observing the poor fit of the Bayesian posterior model, we found a simple heuristic that seem to well capture the explanation preference. Given that people exhibited little sensitivity to the prior manipulation, we hypothesize a heuristic where people only pay attention to the causal strength of potential explanations. Specifically, people:

1. Recognize all the causes that have maximum causal strength.

2. Choose the explanation that includes all the maximum causal strength factors but nothing extra.

For example, when the three causes have causal strength of 0.1, 0.9 and 0.1 in order, then the most preferred explanation is including the first two causes (so that the strongest cause with 0.9 is included) but not more. In another case, where all three causes have equal causal strength, then the best explanation is to include all three causes.

Using this heuristic, we can perfectly predict the most favored explanation, as shown in Table 3.1.

trial	causal strength	predicted best	empirical best
0	A=0.9,B=0.9,C=0.9	A+B+C	A+B+C (64.9%)
1	A=0.9,B=0.9,C=0.9	A+B+C	A+B+C (68.9%)
2	A=0.9,B=0.9,C=0.9	A+B+C	A+B+C (77.0%)
3	A=0.1,B=0.1,C=0.9	A+B+C	A+B+C (85.1%)
4	A=0.9,B=0.9,C=0.1	A+B	A+B (81.1%)
5	A=0.1,B=0.9,C=0.1	A+B	A+B (68.9%)
6	A=0.3,B=0.3,C=0.9	A+B+C	A+B+C (66.2%)
7	A=0.3,B=0.3,C=0.9	A+B+C	A+B+C (78.4%)

Table 3.1: Comparing the predicted best explanation with empirical average best explanation, where 2 means 2-cause explanation, 3-means 3-cause explanation, and so on. The prediction is generated from the heuristic model that chooses the explanation including all the maximum causal strength factors but nothing extra. The empirically best explanation and the proportion of participants choosing this answer are listed in the last column.

3.5.2.3 Heuristic model explains most of the free response data

For each explanation-seeking trial, in addition to the forced choice answers, we included an additional optional question allowing participants to freely combine the causes to indicate the best explanation they have in mind. For example, in all response options the first cause is always included (see section 3.4), but in free response the participant could favor a combination of causes without the first cause. To account for this data, the heuristic model could be adapted by changing the second step to:

2a. Assembling all the factors with strongest causal strength to make that the most preferred explanation.

We can compare the heuristic model predictions for free responses with the predictions from Bayesian posterior model. For the latter, the best explanation is chosen by calculating the posteriors of all the possible combination of causes and pick the best one. In Table 3.2, we can see that the heuristic model predicts the free response data better than the Bayesian model, getting only one out of nine trials predicted wrong while the latter gets 3 trials wrong. Interestingly though, the Bayesian model gets trial 6 right but the heuristic fails to take into account of the prior and thus ignored the cause with a strong prior, giving a wrong prediction. This indicates a significant caveat for the heuristic model that we will address further in Experiment 2.

3.5.2.4 Individual difference in explanation preference

Besides the most popular choices for each problem type, we are also interested in individual difference among participants. For example, when looking into the aggregate choice probability in

trial	prior	causal strength	empirical best	heuristic best	bayes best
0	A=0.1,B=0.1,C=0.1	A=0.9,B=0.9,C=0.9	A, B and C (64.9%)	A, B and C	only A
1	A=0.5,B=0.5,C=0.5	A=0.9,B=0.9,C=0.9	A, B and C (68.9%)	A, B and C	A, B and C
2	A=0.9,B=0.9,C=0.9	A=0.9,B=0.9,C=0.9	A, B and C (77.0%)	A, B and C	A, B and C
3	A=0.3,B=0.3,C=0.3	A=0.1,B=0.1,C=0.9	only C (83.8%)	only C	only C
4	A=0.3,B=0.3,C=0.3	A=0.9,B=0.9,C=0.1	only A and B (79.7%)	only A and B	only A
5	A=0.3,B=0.3,C=0.3	A=0.1,B=0.9,C=0.1	only B (78.4%)	only B	only B
6	A=0.9,B=0.3,C=0.3	A=0.3,B=0.3,C=0.9	only C and A (40.5%)	only C	only C and A
7	A=0.5,B=0.5,C=0.3	A=0.3,B=0.3,C=0.9	only C (39.2%)	only C	only A and B

Table 3.2: Comparing the predicted best explanation from heuristic model and Bayesian model with empirical average best explanation in the free response data.

Figure 3.3, we found a pattern where people choose 1-cause and 3-cause more often than the 2-cause explanation. Is this a general pattern across participants, or does that originate from two sub groups of participants, some always prefer more complex explanations and some others always prefer simplicity?

To answer that, we performed the model-free Agglomerative Clustering algorithm (Pedregosa *et al.*, 2011) which separated all participants into two groups. We then redid the plot for the trials in Figure 3.3, but for the two clusters respectively. In Figure 3.6, we can see that indeed, the aggregate choice ratio for group 1 is almost consistently choosing 3-cause for all the three trials, showing a strong complexity preference. Group 2, on the other hand, strongly prefers the 1-cause i.e., the simplest explanation.

What is the origin of this individual difference? One trivial hypotheses would be that the second group of participants just mindlessly choosing 3-cause at all time. This pattern, however, would not remain for the other trials, especially for those with a cause of strong causal strength – participants will not choose 3-cause in this case. Another possibility is that these participants are purely having very faulty understanding of prior and conjunctive prior, making them ignore that a conjunction of three rare causes should have very low prior probability. Looking into the judgments on conjunctive priors, however, we found that the two groups have shown similar response patterns instead of one group qualitatively different from the other, making this hypothesis less likely (see Figure 3.7). In sum, we found that participants could be grouped into 2 clusters with one cluster strongly prefer complex explanations and the other does not. The underlying mechanism for that difference is still unclear.

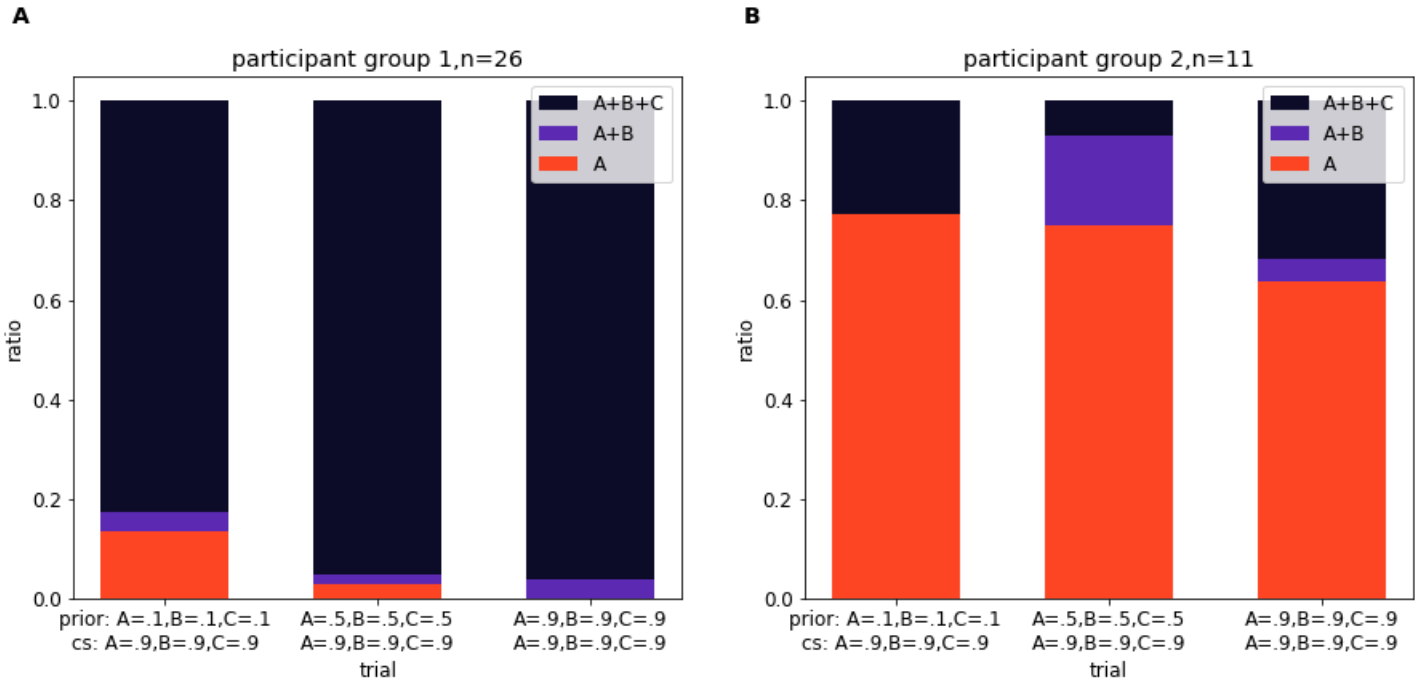

Figure 3.6: Choice ratio of the three explanations for trials with equal prior and causal strengths, separated by the two clusters of participants.

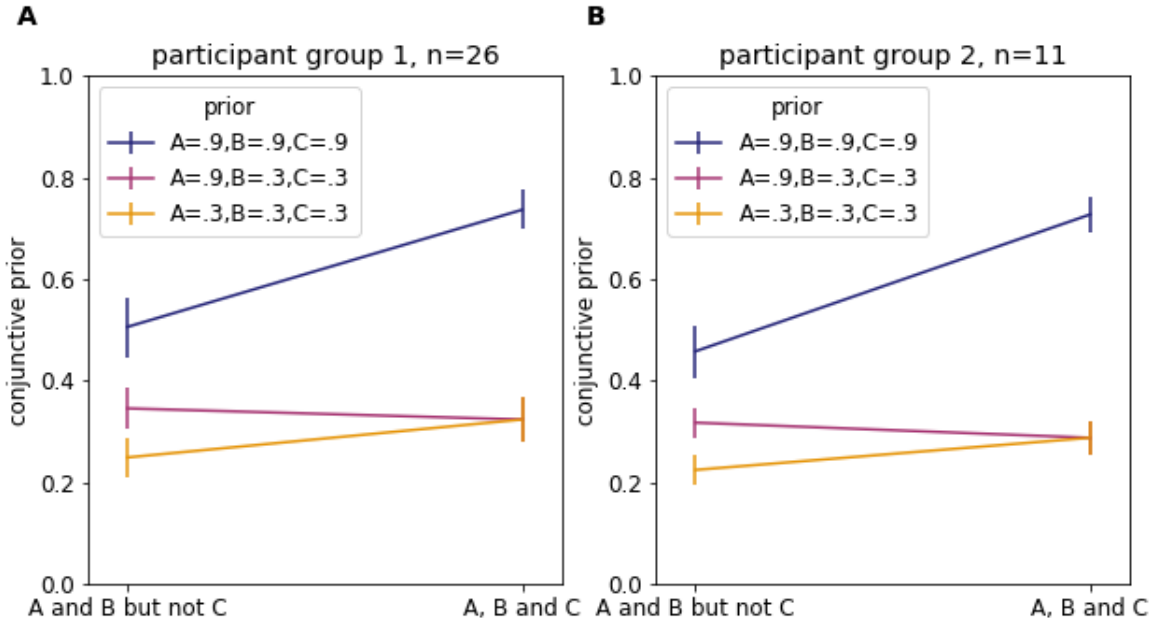

Figure 3.7: Rating of conjunctive priors, separated by the two clusters of participants.

3.5.2.5 Bias in conjunctive prior and causal strength judgments

To understand why the Bayesian model performs so bad, we investigated participants' judgements regarding the prior and causal strength of conjunctive causes. In Figure 3.8 we can see that the overall ranking of participant's average (solid line) for all trials do agree with the theory(dash line), indicating a decent understanding of conjunctive probabilities. The absolute magnitude, however, shows systematic bias with priors being in general higher than the theory and causal strength lower than the theory prediction. Interesting, there seems to be some conceptual misunderstanding prevalent in our data. For example, regarding conjunctive prior, people are significantly over-estimating the prior of 2-cause existing with one other high probability cause being absent. Also, regarding conjunctive causal strength, people seem to think some kind of average instead of adding up algorithm, so that the conjunction of a high causal strength (0.9) cause with lower causal strength cause (0.3) results in something in-between.

We have not found a good framework to explain these deviations from standard Bayesisan formalism. One possibility is to add additional parameters to characterize these deviations of conjunctive prior and likelihood, then feed those into the Bayesian model.

We also have not found an easy way to combine these results to make sense of the explanation preference. But we have excluded the possibility that the origin of individual difference of explanation preference being the differences in conjunctive judgments (see the section above). To put it the other way, these conjunctive judgments cannot fully explain the failure the Bayesian posterior model in explaining the behavioral data.

3.5.3 Discussion

This experiment is a preliminary exploration regarding explanation preference. The new experimental paradigm probing people's preference exposed relatively consistent trends among people, making it a viable tool for exploring this judgment.

We found that for trials with identically probable and strong causes (i.e., equal prior and equal causal strength) in a common effect structure, people, instead of having a simplicity preference as the previous literature suggested (Lombrozo, 2007; Zemla *et al.*, 2020), showed a complexity preference. This behavior is also contradictory to the prediction from the Bayesian posterior model, which also does not provide a satisfying description of the empirical data for the other types of trials (i.e., trials with unequal prior and unequal causal strength).

These data led us to propose a new heuristic model that perfectly explains all the forced choice data as well as most of the free response data. This model prefers the explanation that includes all the causes with maximum causal strength, but not more. This "including all strong causes" may be interpreted as the tendency for complexity preference, whereas "not more" may reflect a simplicity

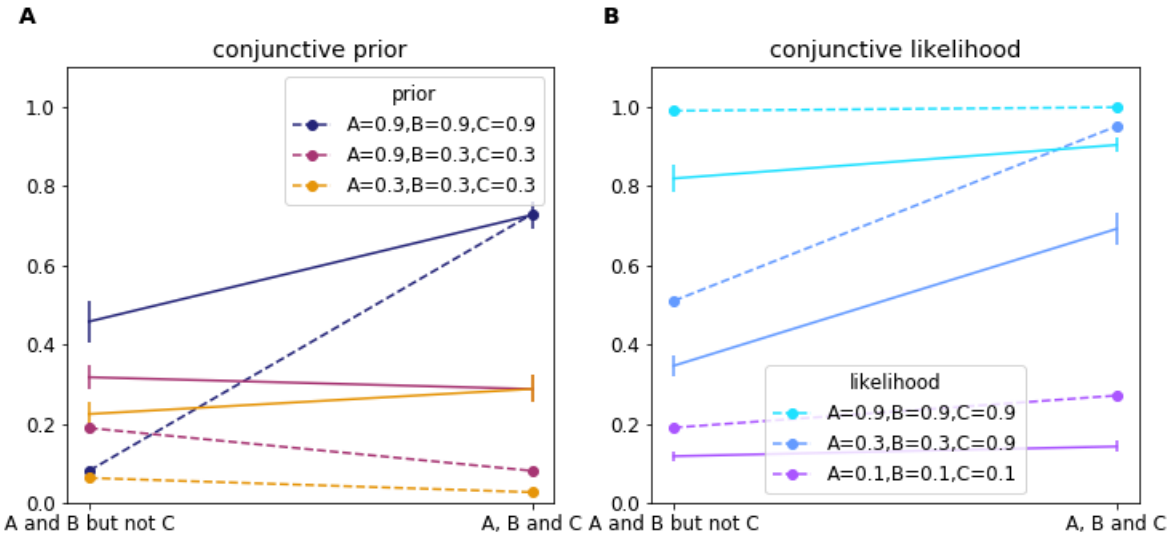

Figure 3.8: Contrasting the empirical data (solid line) with theory prediction (dash line) regarding the prior and likelihood (causal strength) of conjunctive causes. The error bars on solid lines represent standard error. Note that participant reports were in the range of 0-100 and here we normalize it to the range of 0 to 1.

preference. Thus, this heuristic is a combination of complexity and simplicity preference that is sensitive to the specific quality of the explanation's causes.

One caveat of this experiment is that the stimuli space is limited. Specifically, the causal strength values are mostly very extreme (0.9 or 0.1), making the causes more look like deterministic rather than probabilistic. Psychologically, people may treat deterministic causes differently than probabilistic cause, thus it is worth exploring other values of the causal strength.

Furthermore, even though the current heuristic explains most data well, it would be puzzling if people indeed completely ignored the information about causes' prior probability. Alternatively, this could be an artifact of limited stimuli types. In Experiment 2, we designed causes that specifically emphasize the effect of prior and test whether the current heuristic model still gives perfect prediction.

3.6 Experiment 2

Experiment 2 used the exact same paradigm as Experiment 1 except changing the specific trial types (i.e., prior and causal strength of the causes) to further test people's explanation preference. Specifically, we designed different groups of trials that aim to either positively confirm the heuristic we developed in Experiment 1 or to critically challenge it.

3.6.1 Method

3.6.1.1 Stimuli design

We designed three groups of trial types to test different aspects of explanation preference.

The first group of trials all have causes with equal prior and equal causal strength, but the magnitude varies in two different levels respectively (2x2 design). In Experiment 1 we already have these type of trials, with causes' causal strength being 0.9 (i.e., $P(\text{effect} = True|\text{cause} = True, \text{other cause} = False) = 0.9$); whereas here, we allowed the causal strengths to be either 0.25 or 0.75. Details of the trial settings are listed in Figure 3.9. These trials aim to test whether the complexity preference of identical causes that we have seen in Experiment 1 still holds in a wider range of causal parameters.

The second group, in contrast, are designed to test the simplicity preference. These trials all have equal prior but unequal causal strength. Specifically, one cause has higher causal strength than the other causes. According to the heuristics we proposed in Experiment 1, people should show simplicity preference and choose the explanation only including the cause with maximum causal strength and discard the other ones. We also allowed the magnitude of prior and causal strength to change in two different levels. Details of the trial settings are listed in Figure 3.10.

The first two groups are all designed to positively confirm the causal strength heuristic, whereas the third group of trials are designed to challenge this heuristic by including causes with very unequal prior. This is to test whether participants will prefer explanations that include those highly probable causes in a way contradicts to the prediction from the heuristics we developed from Experiment 1.

3.6.1.2 Participants

We recruited 63 participants (average age 35.6, standard deviation 10.4; 34 reported as males and 29 as females) from Amazon Mechanical Turk via Psiturk (Gureckis *et al.*, 2016). We used the exact same exclusion criteria as in Experiment 1 (see section 3.5.1.3). We included in total 28 participants to the final analysis. On average, each participant has 0.64 invalid trials (out of 25) being excluded from the analysis.

3.6.2 Results

As shown in Figure 3.9, for the trials with identical causes, despite both the prior and causal strengths varying, these trial all show explanation preference very similar to Experiment 1 (Figure 3.3). Specifically, in all those trials, the explanation "A+B+C" is the most preferred explanation overall, indicating a form of "complexity preference". But there is also evidence of "simplicity

regressors	Estimate	Est.Error	Q2.5	Q97.5
P(A) Intercept	0.14	0.68	-1.23	1.42
P(A+B+C) Intercept	**2.08**	**0.64**	**0.89**	**3.45**
P(A)_prior	-0.43	0.82	-2.08	1.15
P(A)_cs	**1.66**	**0.86**	**0.00**	**3.40**
P(A)_interaction	-0.79	1.15	-3.11	1.46
P(A+B+C)_prior	0.51	0.68	-0.84	1.82
P(A+B+C)_cs	0.79	0.81	-0.77	2.39
P(A+B+C)_interaction	-1.76	1.05	-3.85	0.32

Table 3.3: Regression results for trials with equal prior and equal causal strength. Dependent variable (DV) are the probability of choosing the explanation "A" or "A+B+C", denoted as P(A) and P(A+B+C), respectively. Intercept is the average baseline of choosing each option, other rows are the fitted slope regarding the specific regressor. "prior" denotes the prior level of high or low, "cs" denotes causal strength level of high or low, "interaction" denotes whether prior and cs are in the same direction or the opposite. The columns represents estimated average, estimated standard error, lower and higher edge of the 95% confidence interval. If the interval includes 0 that means the regressor is not significant.

preference" in that for all trials the explanation "A+B" is less preferred than simply "A". Furthermore, since we have the 2x2 design for these trials, we can then test how the prior and causal strength affects the preference. Figure 3.9B and C summarize the probability of choosing "A" or "A+B+C" given the prior and causal strength condition. We saw that the probability of choosing the 3-cause is significantly higher than the baseline (1/3). None of the condition factors significantly manipulated the choice probability.

To quantitatively examine these patterns, we performed a logistic regression over the probability of choosing each explanation option with a categorical model (the special case of multinomial distribution with only single draw). Regressors include the prior level (high=1, low=0), causal strength level (high=1, low=0), and the interaction of the two. We used the R package brms (Bürkner, 2017), implemented with the probabilistic programming language Stan (Carpenter *et al.*, 2017)) to perform the regression with Bayesian mixed-effect models. The model prior was set to default, i.e., randomly choosing between three options. Table 3.3 shows that the baseline probability of choosing 3-cause is higher than random (mean 2.08, 95% uncertainty interval between 0.89 and 3.45). In the higher causal strength conditions the probability of choosing 1-cause is marginally higher (mean 1.66, 95% uncertainty interval between 0.00 and 3.40). None of the other factors significantly contribute to the explanation preference.

Figure 3.10A, reveals that for trials with one cause having distinctively high causal strength, people are significantly more likely to choose an explanation only including this cause. This is a clear demonstration of simplicity preference. Figure 3.10B indicates that either prior of the

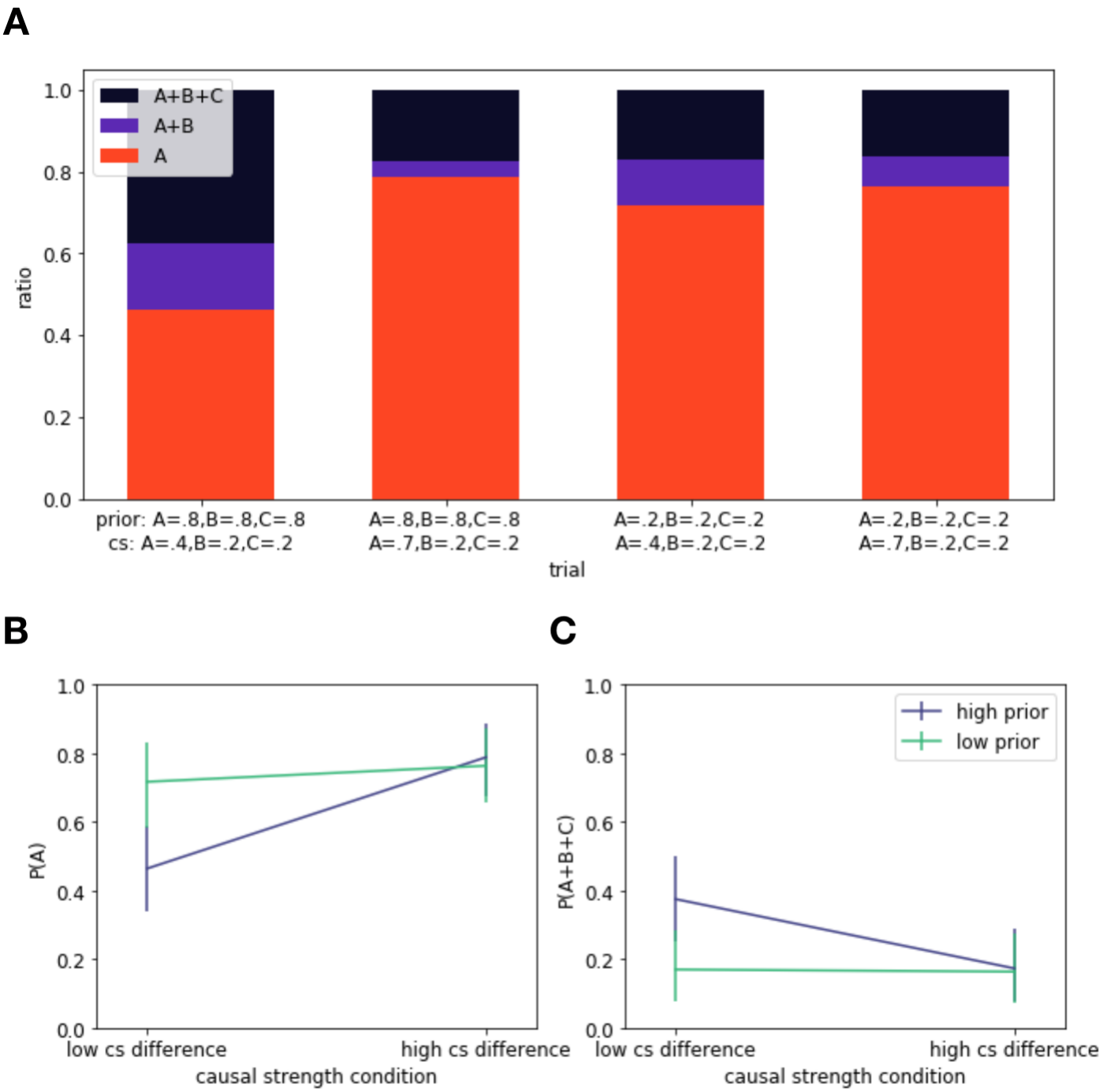

Figure 3.9: Empirical data of trials with equal causal strength and equal prior. A: Averaged ratio of choosing 1-, 2- or 3-cause; error bars indicate the 95% confidence interval calculated from bootstrapping. In the legend, "p" denotes prior and "cs" denotes causal strength of each trial. B: average ratio of choosing simple cause A, separated by the level of prior and causal strength level. C: average ratio of choosing complex cause A+B+C, separated by the level of prior and causal strength level.

regressors	Estimate	Est.Error	Q2.5	Q97.5
P(A) Intercept	2.34	0.65	1.12	3.69
P(A+B+C) Intercept	0.38	0.61	-0.84	1.58
P(A)_prior	-1.27	0.68	-2.66	0.02
P(A)_cs	0.66	0.78	-0.88	2.18
P(A)_interaction	2.16	1.24	-0.13	4.66
P(A+B+C)_prior	0.53	0.70	-0.86	1.86
P(A+B+C)_cs	0.45	0.86	-1.19	2.15
P(A+B+C)_interaction	0.32	1.29	-2.13	2.86

Table 3.4: Regression for the equal prior and unequal causal strength trials. Note "cs" here denotes causal strength difference.

level of causal strength difference does not significantly modulate the choice probability. Again we performed logistic regression on this group of trials, all the same set up as above except the factor of causal strength represents the magnitude of causal strength difference. Only the baseline probability of choosing 1-cause is significantly higher than random and none other factors are significant, as is summarized in Table 3.4.

All the above results are in agreement with the heuristic we proposed in Experiment 1 that people ignore the prior and only taking into account of causal strength to make decision about best explanation. Yet we still need to test that when the prior of various causes have distinctive differences, will people also take that into account. Specifically, to demonstrate the effect of prior, we make the following two groups of comparison:

First, comparing trials of equal causal strength but equal or unequal prior. Figure 3.11 shows that if there is a cause with distinctively high prior, then the most preferred explanation is the smallest number of causes that include this distinctive cause; this is in contrast to the trials with causes of equal prior, where people are more likely to choose 3-cause (a recapitulation of the conclusion in Figure 3.9). Thus the unequal prior changed the complexity preference, making a simpler explanation more preferable. Results from regression analysis quantitatively confirmed this conclusion (see appendix 3.8.2.1).

Second, comparing trials of unequal causal strength but equal or unequal prior. This comparison more shapely tests the heuristic we proposed based on Experiment one, in that if people only pay attention to the causal strength, then people should prefer the 1-cause explanation. However, as shown in Figure 3.12, people do show a strong preference for 2-cause if the second cause has distinctively higher prior (although with lower causal strength). This is a clear evidence that people not only take into account of causal strength but also prior for determining the best explanation.

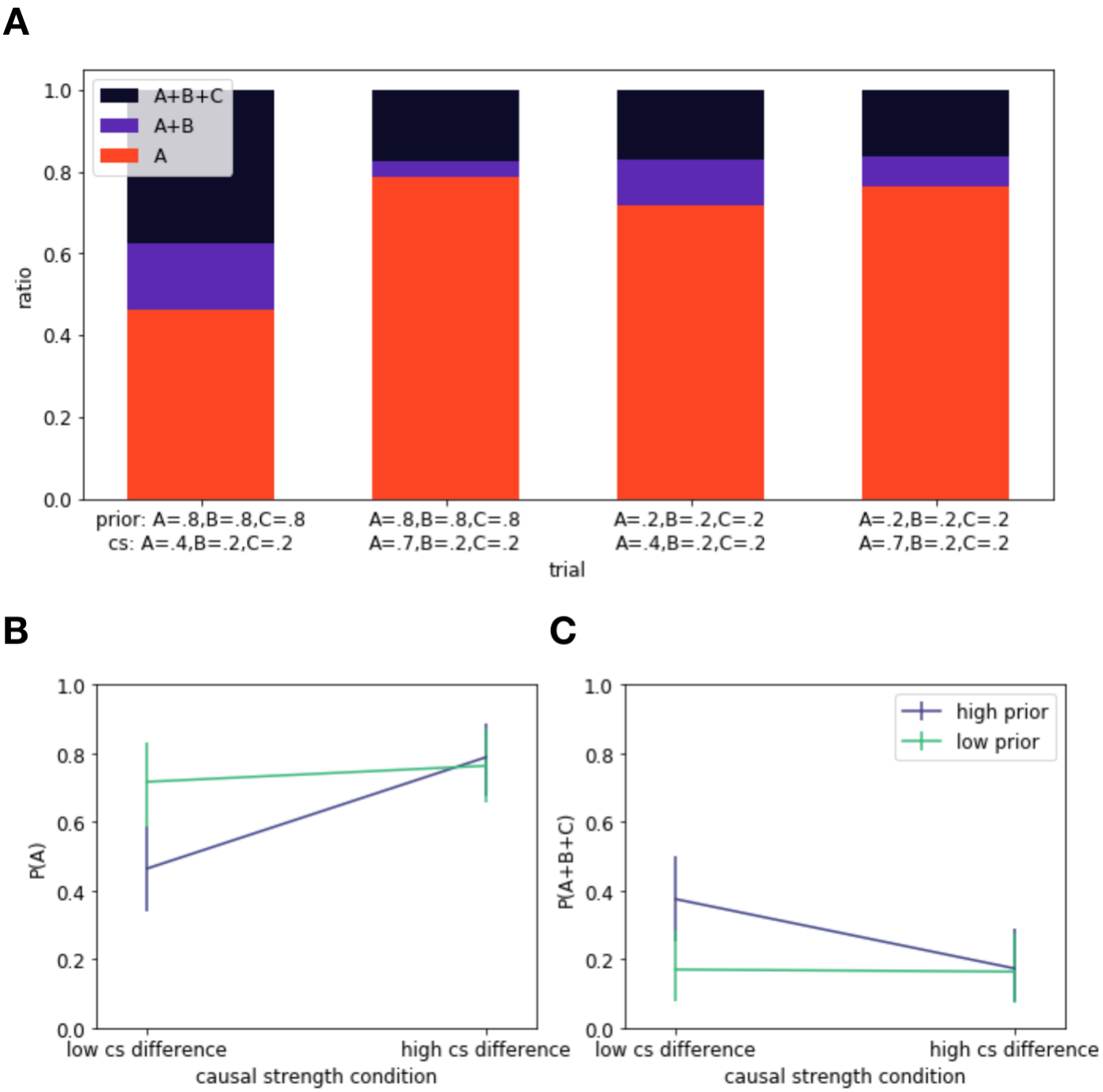

Figure 3.10: Empirical data of trials with unequal causal strength and equal prior. A: Averaged ratio of choosing 1-, 2- or 3-cause; error bars indicate the 95% confidence interval calculated from bootstrapping. In the legend, "p" denotes prior and "cs" denotes causal strength of each trial. B: average ratio of choosing simple cause A, separated by the level of prior and causal strength level. C: average ratio of choosing complex cause A+B+C, separated by the level of prior and causal strength level.

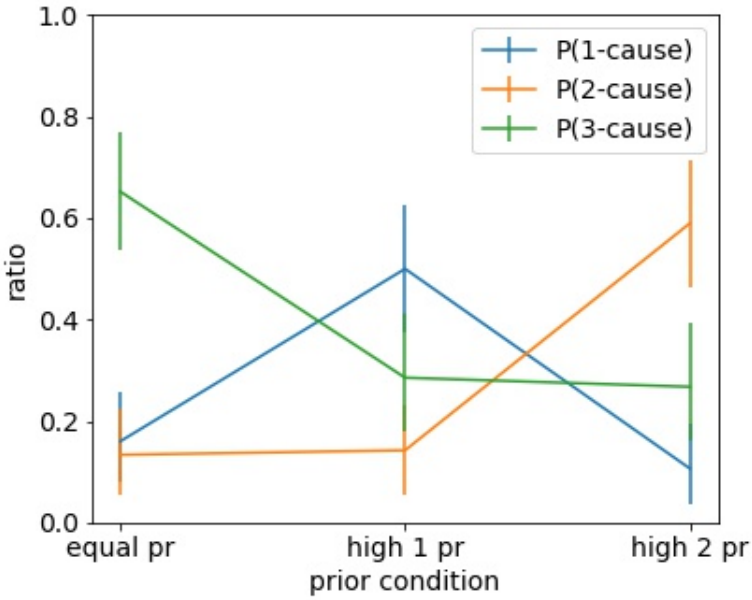

Figure 3.11: Empirical choice probability for trials of equal causal strength but equal or unequal prior. We categorized trials by their prior condition, by whether the trial has equal prior, or the first cause having the highest prior ("high 1 pr" in the figure), or the second having the highest prior ("high 2 pr"). For each trial type, the ratio of choosing one-, two- or three-cause explanations is shown in different colored lines.

3.6.2.1 Heuristic model explains both the forced choice and free response data

Because of the new evidences from unequal prior trials, we developed an updated heuristic to determine the best explanation that takes into account of both causal strength and prior. Thus:

1. Recognize all the causes that either have maximum causal strength or maximum prior. They are all referred to as "distinguishable causes".

2. Choose the explanation that includes all the "distinguishable causes" but nothing extra.

Then based on this heuristic, we can explain all the empirically most preferred explanation perfectly (see Table 3.5 for forced choice data and Table 3.6 for free response data). Note that in the free response table we also present the predictions from the "causal strength only" model. As we can see, trial 8 to 10 were specifically designed to challenge that model and indeed these trials are better explained by the full heuristic rather than the "causal strength only" heuristic.

3.6.2.2 Likelihood fitting

To address the problem that the heuristic model only gives the most preferred explanation instead of a probability distribution of each potential explanation being chosen, we further expanded this

trial	prior	causal strength	predicted best	empirical best
0	A=0.8,B=0.8,C=0.8	A=0.25,B=0.25,C=0.25	A+B+C (73.6%)	A+B+C
1	A=0.8,B=0.8,C=0.8	A=0.75,B=0.75,C=0.75	A+B+C (52.7%)	A+B+C
2	A=0.2,B=0.2,C=0.2	A=0.25,B=0.25,C=0.25	A+B+C (64.2%)	A+B+C
3	A=0.2,B=0.2,C=0.2	A=0.75,B=0.75,C=0.75	A+B+C (57.4%)	A+B+C
4	A=0.8,B=0.8,C=0.8	A=0.4,B=0.2,C=0.2	A (46.4%)	A
5	A=0.8,B=0.8,C=0.8	A=0.7,B=0.2,C=0.2	A (78.8%)	A
6	A=0.2,B=0.2,C=0.2	A=0.4,B=0.2,C=0.2	A (71.7%)	A
7	A=0.2,B=0.2,C=0.2	A=0.7,B=0.2,C=0.2	A (76.4%)	A
8	A=0.1,B=0.8,C=0.1	A=0.25,B=0.25,C=0.25	A+B (61.1%)	A+B
9	A=0.1,B=0.8,C=0.1	A=0.4,B=0.2,C=0.2	A+B (76.4%)	A+B
10	A=0.8,B=0.1,C=0.1	A=0.25,B=0.25,C=0.25	A (53.8%)	A

Table 3.5: Comparing the predicted best explanation with empirical average best explanation. The prediction is from the heuristic that chooses the explanation including all the factors of maximum prior and maximum causal strength but not more than that. The empirically best explanation and the proportion of participants choosing this answer are listed in the last column.

trial	**prior**	**causal strength**	**empirical best**	**heuristic_cs best**	**heuristic_full best**
0	[0.8, 0.8, 0.8]	[0.25, 0.25, 0.25]	A, B and C (71.7%)	A, B and C	A, B and C
1	[0.8, 0.8, 0.8]	[0.75, 0.75, 0.75]	A, B and C (49.1%)	A, B and C	A, B and C
2	[0.2, 0.2, 0.2]	[0.25, 0.25, 0.25]	A, B and C (62.3%)	A, B and C	A, B and C
3	[0.2, 0.2, 0.2]	[0.75, 0.75, 0.75]	A, B and C (51.9%)	A, B and C	A, B and C
4	[0.8, 0.8, 0.8]	[0.4, 0.2, 0.2]	only A (44.6%)	only A	only A
5	[0.8, 0.8, 0.8]	[0.7, 0.2, 0.2]	only A (75.0%)	only A	only A
6	[0.2, 0.2, 0.2]	[0.4, 0.2, 0.2]	only A (71.7%)	only A	only A
7	[0.2, 0.2, 0.2]	[0.7, 0.2, 0.2]	only A (80.0%)	only A	only A
8	[0.1, 0.8, 0.1]	[0.25, 0.25, 0.25]	only B (38.9%)	A, B and C	only B
9	[0.1, 0.8, 0.1]	[0.4, 0.2, 0.2]	only A and B (60.0%)	only A	only A and B
10	[0.8, 0.1, 0.1]	[0.25, 0.25, 0.25]	only A (51.9%)	A, B and C	only A

Table 3.6: Comparing the predicted best explanation with empirical average best explanation for the free response data. "cs heuristic" is the one we proposed in Experiment 1 which only takes into account of causal strength; while the "full heuristic" is the one we presented above. The empirical best comes with the percentage of participants choosing this answer. For the free response, there are in total 7 possible responses therefore the random baseline is 14.8%.

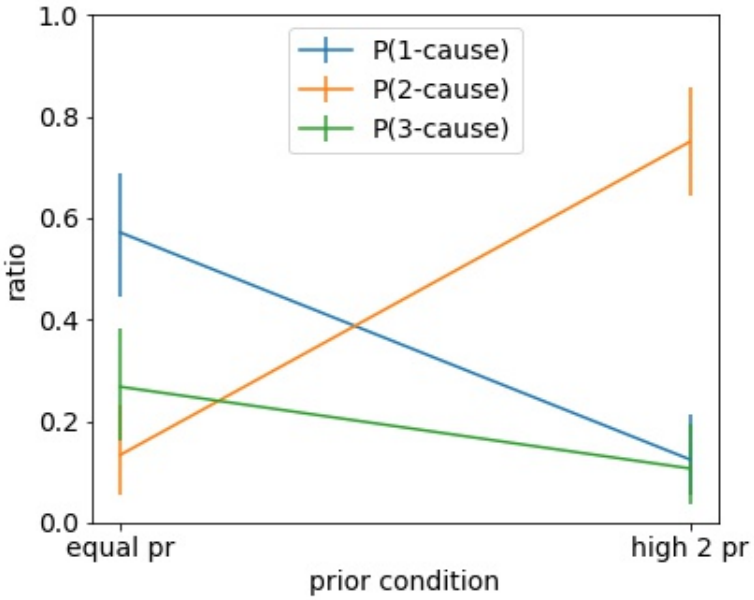

Figure 3.12: Empirical choice probability for trials of unequal causal strength but equal or unequal prior. The first cause always has the highest causal strength. We categorized trials by their prior condition, by whether the trial has equal prior, or the second having the highest prior ("high 2 pr"). For each trial type, the ratio of choosing one- or two- or three-cause explanations is shown in different colored lines.

model with a softmax function that allows some randomness, so that the other explanations can be chosen with a constant probability:

$$P(e_j) = \frac{\exp^{q(e_j)/T}}{\sum_{k=1}^{3} \exp^{q(e_k)/T}} \tag{3.5}$$

where $q(e_j)$ is the heuristic model predicted probability of choosing explanation j (value being 1 or 0 in this case), T is the positive noise parameter (also referred to "temperature") where higher T is associates with more noise, i.e., more equal probability for each explanation despite the model predicts otherwise. Alternatively if $T \to 0$ then whichever explanation is recommended by the heuristic model will also output $P(e_j) = 1$ and the other options have no chance to be chosen. Similarly we can add the same softmax noise for Bayesian posterior model to make it adapt to different individuals.

We fit each participant individually for the noise parameter T and calculated the maximum likelihood averaged across trials. In Figure3.13, we show the likelihood difference of heuristic model and Bayesian model. An overwhelming number of participants get better fit from the heuristic model.

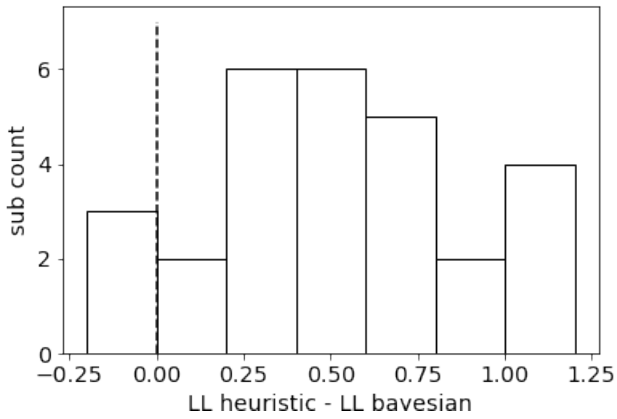

Figure 3.13: The likelihood difference between heuristic model and Bayesian model for each participant.

trial	prior	causal strength	empirical best	heuristic_cs best	heuristic_full best
0	[0.1, 0.1, 0.1]	[0.9, 0.9, 0.9]	A, B and C (64.9%)	A, B and C	A, B and C
1	[0.5, 0.5, 0.5]	[0.9, 0.9, 0.9]	A, B and C (68.9%)	A, B and C	A, B and C
2	[0.9, 0.9, 0.9]	[0.9, 0.9, 0.9]	A, B and C (77.0%)	A, B and C	A, B and C
3	[0.3, 0.3, 0.3]	[0.1, 0.1, 0.9]	only C (83.8%)	only C	only C
4	[0.3, 0.3, 0.3]	[0.9, 0.9, 0.1]	only A and B (79.7%)	only A and B	only A and B
5	[0.3, 0.3, 0.3]	[0.1, 0.9, 0.1]	only B (78.4%)	only B	only B
6	[0.9, 0.3, 0.3]	[0.3, 0.3, 0.9]	only C and A (40.5%)	only C	only C and A
7	[0.5, 0.5, 0.3]	[0.3, 0.3, 0.9]	only C (39.2%)	only C	A, B and C

Table 3.7: Comparing the predicted best explanation with empirical average best explanation for data in Experiment 1. The "cs heuristic" is the one we proposed in Experiment 1 while the "full heuristic" is the one we presented in the current section.

3.6.2.3 Explaining data in Experiment 1 with the new heuristic

Now we have established the effectiveness of the full heuristic model, one additional check is to apply it to the data from Experiment 1. As shown in Table 3.7, for most trials the two heuristics generate same prediction except for trial 6 where the full heuristic is correct by taking into account of the high prior cause. Trial 7, on the other hand, is the only trial that the full heuristic fails to explain the most preferred cause where participants seem to have ignored the high prior trials since the prior difference (0.5 or 0.3) is not as big. This is a point for future discussion of the limitation for our current heuristic model.

3.6.3 Discussion

In Experiment 2 we expanded the range of stimuli variety compared to Experiment 1. Current data showed that rather than the absolute magnitude of the prior or causal strength, it is the relative magnitude between relevant causes that drives the explanation preference. As is shown in the regression analysis, the prior of causal strength factors are not significant contributors to the probability of choosing simple or complex explanations. This lack of significance could of course be due to inadequate amount of data, or the stimuli value manipulation being not informative enough. Future research could potentially give more decisive answers on this issue.

On the theoretical side, we have developed a full heuristic model that takes into account of both the prior and causal strength information. This model gives almost perfect explanation for both the forced choice and free response data in terms of most preferred explanation. With an expansion of softmax noise, it could fit the individual subject data much better than the Bayesian posterior model. Moreover, it could almost perfectly cover the data from Experiment 1, except for one trial, where the predicted best explanation from this model is more complex than the empirical best. With denser sampling of the stimulus parameters, it is worth exploring whether people have further rules that makes a simpler explanation more preferable.

In sum, our new behavioral paradigm has demonstrated novel behavior patterns, which cannot be explained by either simplicity / complexity preference or Bayesian posterior model. The heuristic of "choosing the explanation with all the distinctively high prior and causal strength causes, but not more than that" is a promising theory for explanation preference.

3.7 General discussion

We used a novel quantitative paradigm to study people's preference for certain explanations. Rather than only testing for simplicity or complexity preference, we found that by manipulating the probabilistic qualities of the causes, we could induce people to prefer different combination of the causes. This procedure illuminates the more fundamental factors underlying the simplicity or complexity virtues of the explanation.

We also tested quantitative models for explanation. The posterior-based model does a worse job of predicting the behavioral data compared to a heuristic model, where the best explanation contains all the causes with distinctively high prior or causal strength, but not more than that.

Our findings add new perspective to the previous literature. First, although a simplicity preference has been validated in many previous literature in empirical studies for both adult (Lombrozo, 2007) and children (Bonawitz & Lombrozo, 2012), it was not a very strong tendency in our paradigm. People prefer simple explanation only when one cause is much more probable (i.e.,

higher prior probability) and / or much more stronger (i.e., higher likelihood for the explanandum) than the other. Otherwise, if more than one causes are distinctively high in either its prior or causal strength, people would tend to include those into a more complex explanation. This could even be seen as an instance of conjunction fallacy Tversky & Kahneman (1983) since this conjunctive explanation could have lower prior and even lower posterior, thus seems like a "fallacy".

In fact, our heuristic model could serve as an alternative explanation to the results in Zemla *et al.* 2020. They found that when comparing a simple or complex (conjunction of two causes) explanation, people prefer the simple explanation if no additional mechanism information is provided. However, note that the simple cause is assigned with a higher causal strength and medium prevalence, while the complex causes each has high prevalence but medium prevalence. They found that the majority of participants actually chose no preference for either one, which agrees with the heuristic because the simple explanation is distinctive in terms of causal strength and the complex is distinctive in terms of prior. Our theory even predicts that if another option of all those three causes are presented as a conjunction, it is going to be the most favored explanation.

Another previously identified phenomena that we did not see in our experiments is from Johnson *et al.* (2019) where they found that when causal effects become more stochastic, i.e., having lower causal strengths, the preference for complex explanation is stronger. In our study, although we do find strong complexity preference where causes with identical prior and causal strengths are likely to be combined together to explain the common effect, this preference is not strongly manipulated by the magnitude of the causal strength (see Table 3.3 and 3.4 where the factor of causal strength is not statistically significant). More evidence is needed to settle on this issue.

There are several potential future extensions based on our study. Regarding the empirical method, our stimuli is limited to the common effect (collider) causal structure with maximum three causes. It is worth studying whether the behavioral pattern would hold if the total number of causes increases; or if the causal structure becomes common cause so that adding more causes will not necessarily increase the likelihood of the effect being existent (as discussed in Zemla *et al.*, 2020). Moreover, our paradigm presents the statistical information in terms of numbers and graphs, yet this may not be a common way for people to learn about causal information. Previous studies have used more experiential way to sequentially present the co-existence rate of causes (Lombrozo, 2007; Pacer & Lombrozo, 2017) or between cause and effects (see a review in Lu *et al.*, 2008). In principle, our paradigm could be adapted to this way of presentation and answer similar questions. Regarding the theory side, our model, even though a coherent rule for generating the best explanation, is also limited to the current paradigm due to a lack of more general computational principle. Future study after gaining more empirical evidence will potentially find the more fundamental mechanism underlying this heuristics (thus, a simpler explanation). Furthermore, this model is not able to explain the individual difference that we have shown in Section 3.5.2.4 since

it does not have any flexible parameters to allow that, which is another future improvement could be done.

3.8 Appendix

3.8.1 Details of the experimental material

3.8.1.1 Transcript of the tutorial

Below is the transcript of the tutorial. The narration is processed with Audacity software to make it sounds like an alien voice.

> Welcome to planet Omega! I am doctor Luzeka. Thank you for agreeing to be my medical assistant here. We are busy dealing with a lot of patients on planet Omega. I would like you to give some judgments on some patient cases about what might be the cause of the symptom. Don't worry about not having the medical knowledge of alien patients: I will give you the relevant information. all you need to do is read the documents and then use your own judgment.
>
> The medical cases look like this:
>
> [show one interface, read the intro]
>
> "Many factors contribute to the symptom of Ozipod pain: Exposure to chemicals such as Wluxia, Metherine and Zithna all lead to Ozipod pain, independently."
>
> Let me stop here and explain the last sentence. In most of our medical cases, the symptom could have several possible explanations. These explanations almost always arise in no relation with others. To use the earth analogy, if you see Emily sneezing a lot, the explanation could be 1) Emily has a cold; 2) Emily has some seasonal allergies; 3) Emily was just cutting a lot of chilli in the kitchen. All these explanations are not related to each other. If a person has a cold, it does not increase or decrease the chance of them having allergies.
>
> That being said, different explanations have different chances to be present. For example, if it is winter right now, then seasonal allergy could be relatively rare, like, among 100 people maybe only 10 would have a seasonal allergy; or if in Emily's culture people very often cook chilli pepper for food, then maybe among 100 random people, 90 will likely to cut a lot of chilli peppers.
>
> Another point to make clear is that all the explanations can cause the symptom on their own. That is, it's not the case that both of cold or allergies have to be present

for sneezing to happen. Rather, having any one of the causes can already make Emily sneeze. But of course, if Emily both catching a cold and having seasonal allergies, then the chance of sneezing will increase.That being said, different explanations could be a strong or weak cause to the symptom. For example, a seasonal allergy could be a relatively strong cause of sneezing, like, among 100 people who have a seasonal allergy, maybe 90 would be also sneezing a lot; at the same time, cutting chilli peppers for maybe not so often causing sneezing (especially if the pepper in that region is not so spicy), meaning that among 100 people who cut the pepper, 10 will be sneezing.

Now let's come back to planet Omega. To prepare you for the task, we will first show you some public health reports from the planet Omega and ask you two kinds of questions: first, given the existence of several different medical conditions, how frequent is it for them to happen at the same time? We will provide you the statistics of each condition in those information cards. You will answer that in the form of among 100 people, how many of them may have certain condition or combination of conditions. You will slide the slider to give your response. Second, given the possibility of several independent medical conditions causing a symptom, how likely the combination of those conditions will cause this symptom? Again you will be given the statistics about each single medical condition, and you will use the slider to report your judgement regarding the combination of those conditions.

After you've finished these questions, you are more familiar with how things work in our planet and we will show you medical documents of individual patients.

In the medical documents for you, the research results from public health department are also included, so you can go ahead click on the documents for each potential explanation and check their data.

Then the final step is to make your final judgments on three potential diagnosis made by other medical assistants. Please please make responsible judgments for these patients. Their symptoms are indeed severe and need diagnosis as accurate as possible. Thank you for helping with this task! Now let's go to do the real works.

The full tutorial video can be downloaded in OSF: https://osf.io/7zbek/.

3.8.1.2 Trials for prior and causal strength judgment

After the tutorial, the first section of the task is about conjunctive prior probability judgement. The introduction for this section says:

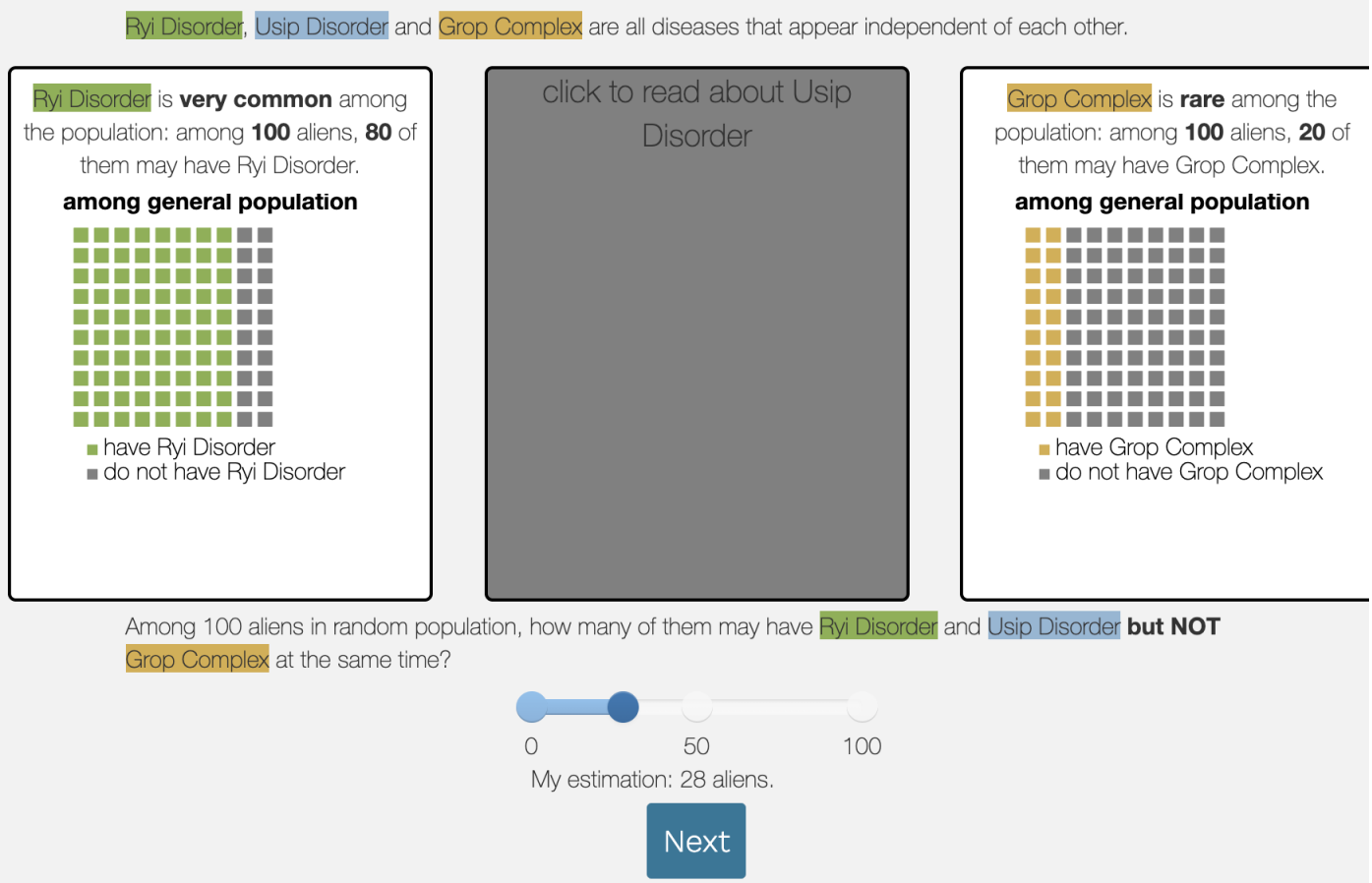

Figure 3.14: An example trial for the prior judgement section.

> Now please make some judgements regarding how often some medical conditions occur.

The interface for these trials are shown in Figure 3.14.

For the second section, participants are asked to judge for conjunctive causal strength. Below is the introduction for this section:

> Now please make some judgements regarding the effects of the combination of several medical conditions.
>
> Remember that most times one symptom may have several causes at the same time. That is, it's not the case that all of these medical conditions have to be present for the symptom to be present. Rather, each condition can independently produce the symptom on its own.

The interface for these trials are shown in Figure 3.15.

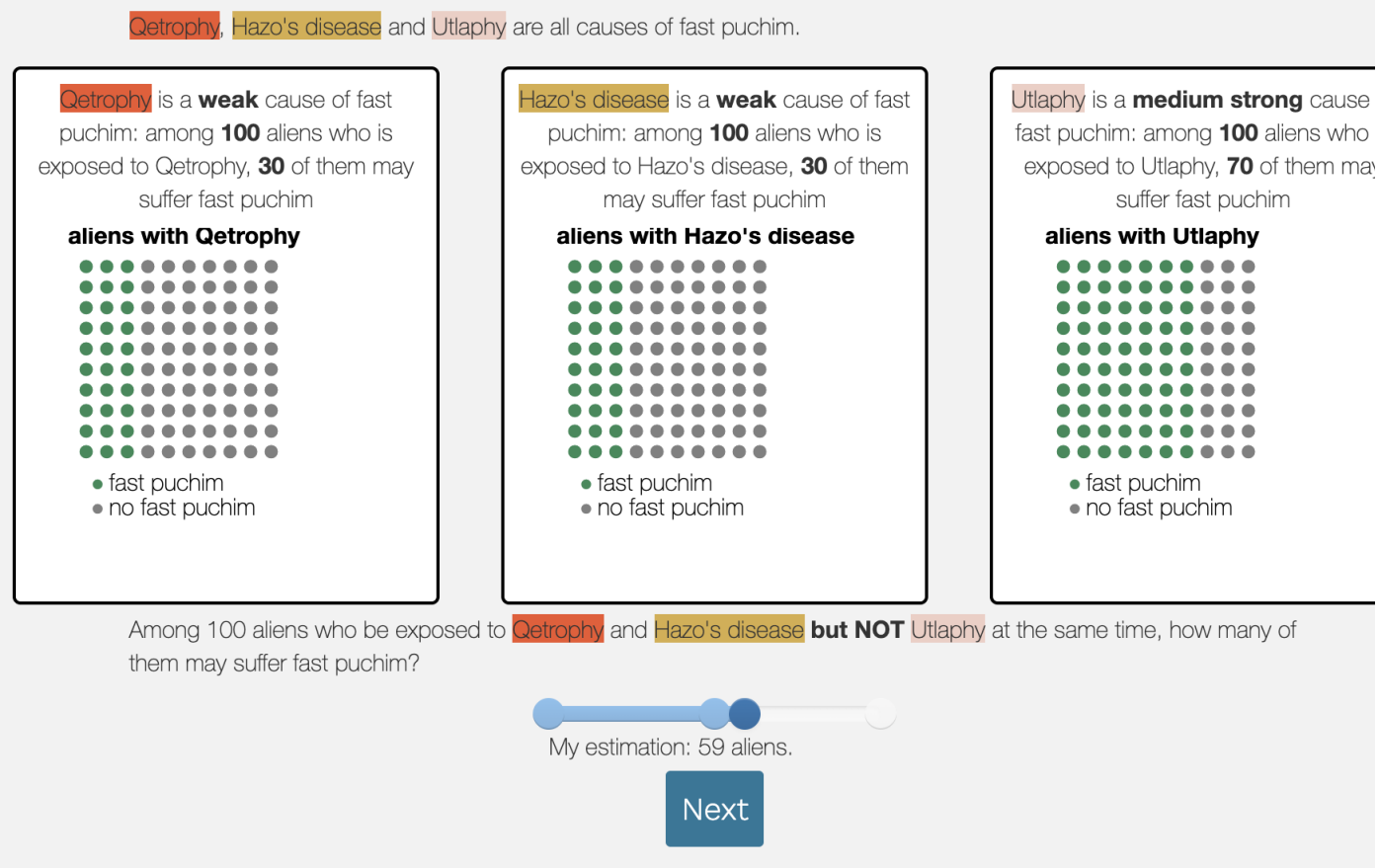

Figure 3.15: An example trial for the causal strength judgement section.

regressors	Estimate	Est.Error	Q2.5	Q97.5
P(A) Intercept	0.07	0.45	-0.82	0.96
P(A+B+C) Intercept	1.91	0.43	1.11	2.79
P(A)_distinct prior 1	1.33	0.58	0.19	2.49
P(A)_distinct prior 2	-2.10	0.64	-3.38	-0.94
P(A+B+C)_distinct prior 1	-1.24	0.59	-2.41	-0.08
P(A+B+C)_distinct prior 2	-2.94	0.52	-4.01	-1.93

Table 3.8: Regression analysis on trials with distinctive priors and equal causal strengths. The distinctive prior could be either the first cause or the second cause given our stimuli design.

3.8.2 Additional results

3.8.2.1 Experiment 2: regression analysis

For trials with distinctive priors, we performed the regression analysis to test whether the prior does significantly bias the explanation preference. The analysis is grouped into two types of trials: equal causal strength and unequal causal strength.

For the equal causal strength trials, we set the regressors to be either the trial has a distinctive prior for the first cause ("distinct prior 1"), or the second cause ("distinct prior 2"), both variables being Boolean. Results in Table 3.8 shows that even though for the baseline condition where all priors are equal, the probability of choosing 3-cause is high, but if the first prior is distinctively high, the probability of choosing 3-cause decreases significantly while the probability of choosing 1-cause increases significantly; if the second prior is distinctively high, both the probability of choosing 1- or 3-cause decreases significantly, indicating people are much more likely to choose the 2-cause explanation (because the probability should sum to 1). For graphic presentation of this result, see Figure 3.11 in the main text.

For the unequal causal strength trials where the first cause always has the highest causal strength, we performed similar regression analysis except that 1) the independent variable is only whether the second cause has a distinctively high prior or not, and 2) the dependent variable is set to be P(chooose 1-cause) and P(choose 2-cause) to more intuitively demonstrate the effect. In Table 3.9 we found that when the causes all have equal prior, the probability of choosing 1-cause is significantly higher than the random baseline and the probability of choosing 2-cause is significantly lower. However, if the second cause has distinctively higher prior, then the probabilty of choosing 2-cause becomes significantly higher. For graphic presentation of this result, see Figure 3.12 in the main text.

regressors	Estimate	Est.Error	Q2.5	Q97.5
P(A) Intercept	0.92	0.38	0.22	1.69
P(A+B) Intercept	-0.98	0.48	-2.05	-0.15
P(A)_distinct prior 2	-1.01	0.74	-2.53	0.42
P(A+B)_distinct prior 2	3.38	0.74	2.05	4.97

Table 3.9: Regression analysis on trials with distinctive priors and unequal causal strengths where the first cause always has the highest causal strength. The trials either have equal priors for all three causes, or the second cause being distinctively high, given our stimuli design.

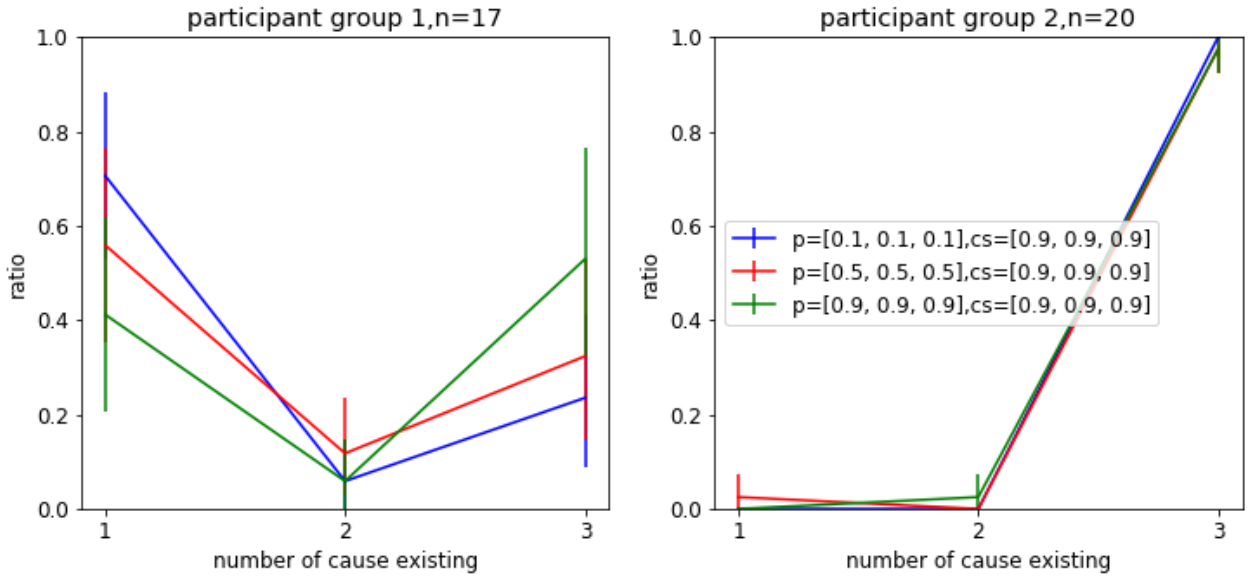

Figure 3.16: Choice ratio of the three explanations, separated by the two clusters of participants.

3.8.2.2 Experiment 2: individual differences

Similiar to the analysis in section 3.5.2.4 in the main text, here we analyze the individual difference of explanation preference for the data in experiment 2 and explore the potential origin of that difference. Specifically, for the trials that have causes with equal causal strength and equal prior (4 trial types, each repeated 2 times, thus 8 trials in total per participant), we performed the model-free Agglomerative Clustering algorithm (cite python sklearn) on those responses which separates all participants into two groups. We then plot the aggregate choice probability for those trials for each group. As is shown in Figure 3.16, the aggregate choice ratio for group 2 is almost consistently choosing 3-cause for all the three trials, showing a strong complexity preference, where as the first group shows slightly more simplicity preference.

To test the origin of this individual difference, again we checked whether the participants in group 2 have very faulty understanding of prior and conjunctive prior, making them ignore that

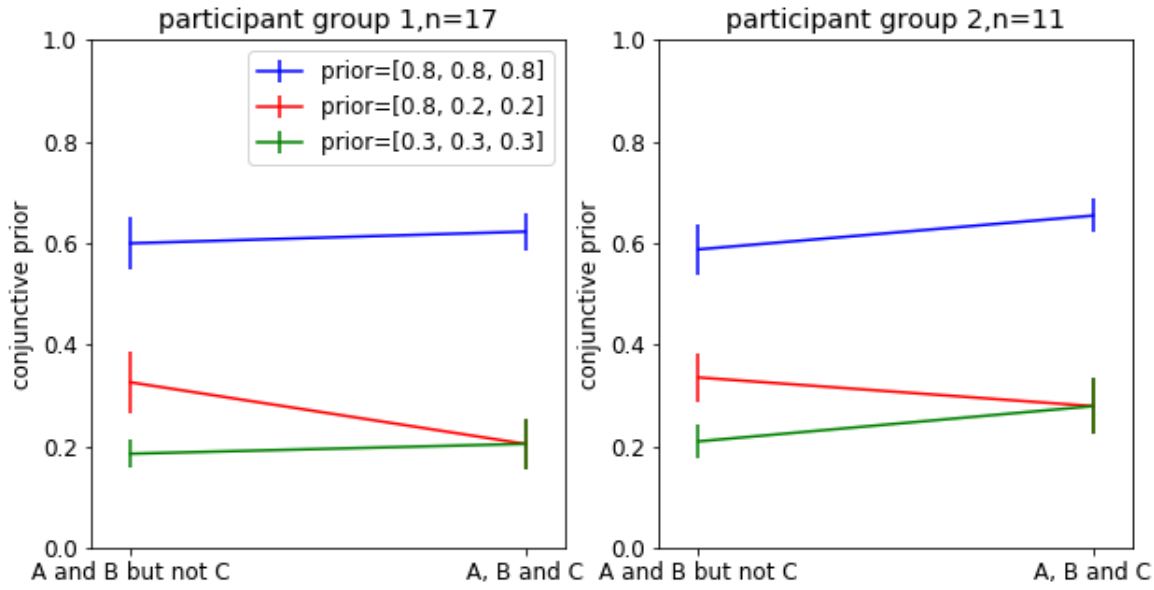

Figure 3.17: Rating of conjunctive priors, separated by the two clusters of participants, for Experiment 2.

a conjunction of three rare causes should have very low prior probability. Examining into the judgments on conjunctive priors, however, we found that, as in Experiment 1, the two groups have shown similar response patterns instead of one group qualitatively different from the other, making this hypothesis less likely (see Figure 3.17), making this potential explanation unlikely.

3.8.2.3 Experiment 2: prior and causal strength judgments

In Experiment 2 we also performed the check of conjunctive prior and causal strength rating. The stimuli was qualitatively similar to Experiment 1 except slight number changes. The patterns in Figure **??** is again very similar to Experiment 1 (Figure 3.8 in the main text), showing systematic deviation from the standard probabilistic theory.

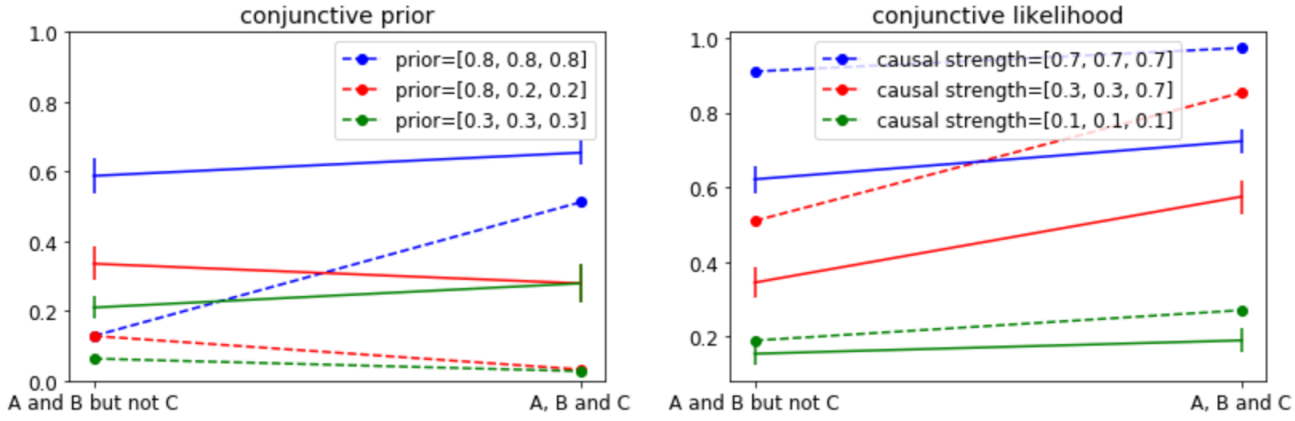

Figure 3.18: Contrasting the empirical data with theory prediction (dash line) regarding the prior and likelihood (causal strength) of conjunctive causes. The error bars on solid lines represent standard error. Note that participant reports were in the range of 0-100 and here we normalize it to the range of 0 to 1.