Probabilistic models of subjective judgments

by

Zhiwei Li

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Center for Neural Science
New York University
2021

Advisor:

Todd Gureckis

Zhiwei Li
zhiwei.li@nyu.edu

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

ABSTRACT

Probabilistic models have been very influential in many cognitive science topics such as language, concepts learning, decision-making etc. The topic of subjective feelings and judgment, however, has been less studied in this computational framework. My thesis focuses on the subjective judgments or emotions that do not obviously relate to instrumental values. My work explores different ways to extract latent variables from the underlying probabilistic model aiming to explain three different subjective judgments: the subjective feeling of suspense, satisfaction of explanation, as well as preference for food.

The first chapter applies a probabilistic model to explain the dynamically changing feeling of suspense preceding the arrival of new information. The central idea of this project is to test and evaluate a model taken from the economic literature (Ely et al., 2015) using a novel empirical paradigm. Succinctly, the model quantifies suspense as the expected belief update in the nearest future, with belief update being quantified as changes in posterior probability. Other heuristics proposed by the larger literature regarding suspense in story telling or movie watching are formalized and compared. Evidence from a variety of stimuli and carefully contrasted conditions indicates that the "future belief update" model best captures the subjective report of suspense.

The second chapter focuses on the feeling of satisfaction of explanation, which is similarly an emotion closely related with the arrival and interpretation of new information. In previous literature, the "simplicity preference" of an explanation 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 and causal strength 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 the prior of each cause existing is low and the causal strength is high. Moreover, a standard Bayesian estimation of the posterior of some explanation being true is not an accurate account of people's preference; rather, people heavily overweight the importance of causal strength than the prior when comparing candidate explanations.

The third and last chapter is a theoretical model regarding how the information collected through active attention relates to how people make value-based decisions. This work is based on the empirical findings from Krajbich et al. that when one snack item has been fixated on more than the other one, it's more likely to be chosen. To explain this effect, a novel model was devel-

oped where the utility of an item is a weighted sum of the posterior mean and the negative posterior standard deviation, with the latter accounting for risk aversion. This model explains the data better than the original attentional drift-diffusion model proposed by Krajbich et al. but worse than a variant with a collapsing bound.

In summary, by constructing different latent variables based upon probabilitic models and testing with new quantitative experiments, my work advances a formal, predictive account of what are previously through to be highly subjective, individualized judgment and emotions.

CHAPTER 1

Introduction

> This game is really fun and engaging. Thank you so much!
>
> — one online participant.
>
> I found this task really long and repetitive. BORING.
>
> — another online participant.

I remember attending my first ever psychology experiment where I was locked in a quiet dark room staring at lines and dots for one hour: the feeling of boredom and drowsiness totally overweighted my interest in psychophysics. The fact is, regardless of what topics the cognitive science experiments are studying (e.g., categorization, decision-making, language learning), the subjective experience induced by participating these experiments is always far wider and richer than those research topics typically allow for. Similarly, when I am reading a scientific paper, what I gain is never simply new knowledge but also a myriad of feelings, such as surprise, amazement, skepticism, dissatisfaction or even amusement. Emotions, feelings, moods are flimsy and elusive affective states that are rarely studied in cognitive science, yet we experience them in almost every conscious moment in life.

Why do we have these affective states? The function of affective states can be broadly categorized into intrapsychic role and interpersonal role (Vallverdú & Giannoccaro, 2015). Interpersonal roles are more about using emotional expressions as a communication signal for achieving relational goals such as showing acceptance or asserting dominance. Intrapsychic roles are more about the individual's process, including not only the homeostasis and survival instincts (e.g. fast recognition of a dangerous stimuli like a tiger), but also many seemingly "cognitive" processes. Contrary to the notion that emotion is the hindrance to rationality, some evidence has shown that emotion is crucial for better cognitive function.

For example, Bechara & Damasio (2005) proposed the "somatic marker hypothesis" stating that rational decision-making is not only about intelligence, but rather critically depends on normal emotional processing. They found that when learning about gambling options, not only the

conscious knowledge of the choice outcome, but also the physiological cues such as increased skin conductance rate, are essential predictors for whether the participant can learn about rewards and punishments then make more advantageous choices later on. Patients with impairments in amygdala or ventromedial (VM) prefrontal cortex have trouble learning from previous experiences and making beneficial personal and social decisions, despite them maintaining a normal problem-solving ability in standard laboratory tests.

Curiosity, as another example, is not only an intrinsic motivation for inquiring more knowledge, but has also been shown to play a role in memory of new information. Kang *et al.* (2009) found that when considering trivia questions, if an question incites more curiosity, people perform better in surprise recall test regarding this question after 1 to 2 weeks. Neural imaging also confirms that curiosity increases activity in memory areas if people guessed incorrectly, indicating that curiosity helps enhancing the encoding for surprising new information. Furthermore, when in a curious state, the consolidation of information is also enhanced through dopaminergic neuromodulation of the hippocampus (Gruber & Ranganath, 2019).

Despite the contribution of affective states to different cognitive processes, it is difficult to study. The most thoroughly studied topic might be fear given the integration of animal model, behavioral paradigms and measures, neural circuit studies. Emotion is generally signified by multiple modalities, including the physiological correlates, neural correlates in ANS and CNS, facial expressions and cognitive bias, among other aspects. Thus it is very hard to use any single aspect to define any specific emotion.

In my work I used two major components to define an affective state: self-reported judgments as the behavioral signature, and probabilistic models as the explanatory framework.

Self-reported judgments are the necessary research component if the focus of study is conscious affective experience. It cannot be replaced by biological measurements of seemingly related states. Again, using the example of fear, LeDoux (2014) has warned that it is important to distinguish the conditioning response to threats and the conscious feeling of fear. Damage to the hippocampus in humans has been shown to disrupt explicit conscious memory of having been conditioned but not the biological conditioning response itself. Some may argue that the self-reported affective state is by definition very subjective thus not a good subject for empirical study. I would say that this subjectiveness is an inevitable consequence if emotional responses are conceptualized as not only the reaction to external input, but also an integration of autobiographical knowledge and introspection (LeDoux & Brown, 2017).

Reports being subjective does not mean they cannot be explained. The classic explanation for emotion is the appraisal theory where people evaluate the current state in different dimensions. For example, Scherer (2001) proposed a 30 dimension model with four broad categories: relevance, implication, coping potential and normative significance.

My work takes a different approach based upon probability. The fundamental assumption is that the brain, rather than representing the world in a certain and deterministic manner, maintains a probabilistic model of the world and makes inference to construct the representation (Knill & Richards, 1996). The brain is also constantly making predictions and simulations regarding the future, comparing with the reality which then guides the future exploration (Friston & Stephan, 2007). These ideas have been widely applied in areas like perception (Walker *et al.*, 2020) as well as language learning (Armeni *et al.*, 2017), physical reasoning (Battaglia *et al.*, 2013), etc.

Despite probabilistic modeling being used for more efficient representation of the external world, less common have the they been associated with subjective states. Here I propose that probabilistic representations are useful in two ways. First, we not only need to have knowledge about external world but it is equally important to represent and have access to our own internal states. When looking at the pictures on a menu to decide on the order, what's involved is not only the external visual stimuli but also my subjective memory with certain kinds of food ("Last few times I ate fish I liked them a lot") and my evaluation of my current physiological state ("Do I want to eat hot food or cold?"). In chapter 3 I will provide an example of building utility function for decision-making that is based on this kind of internal representation of value. Second, many internal states are indicators for how much attention one need to pay to certain information source, how much efforts one should make for collecting more information — in sum, these affective states serves as the regulator for learning (von Haugwitz & Dodig-Crnkovic, 2015). For example, the feeling of boredom in doing psychophysics experiments as I mentioned at the beginning of the introduction, can be seen as an informatory cue drawing my attention to the unsatisfactory state I am in, as well as an motivator towards some other more meaningful and satisfying tasks (Elpidorou, 2018). In the first two chapters I will focus on two affective states that I see as part of our information regulation mechanisms: the feeling of suspense in face of information to be revealed (Chapter 1); the satisfaction to an explanation given the statistical information of the causal structure (Chapter 2).

In a broader picture, by introducing the probabilistic modeling and novel behavioral paradigms for quantitative studying of affective states, I hope my work provides new perspectives on these topics. I also hope this work could call attention for more researchers from the computational modeling background to use their expertise for exploring wider span of topics that, despite elusive, are quantifiable and essential for understanding human conscious experience.

1.1 Modeling suspense as expected future learning

When does a sport match become most suspenseful, that the audience has to hold their breath, forget about eating popcorn or going to the bathroom, paying full attention to the game so they

do not miss anything? Usually this does not happen at the very beginning of the game because whichever team wins a point is not very consequential. Towards the end of the game, it may still not be suspenseful if one team already have a big advantage that the other side has no chance to flip the situation. However, if the game is towards the end and both sides have a fair chance of winning, then the game could be come really intense and suspenseful. Is this a universal intuition that people would generally agree with, regarding their feeling of suspense? If so, what would be a good way to explain this mechanism?

Empirically, previous studies showed that people do have general agreements to feel more suspense in certain conditions. For example, in a story-telling setting, people may feel more suspense if the chance of the protagonist fails is high and the possible solution for the protagonist has been removed (Comisky & Bryant, 1982; Gerrig & Bernardo, 1994); also, the presence of time pressure could also increases suspense (Alwitt, 2002). What could be an underlying principles behind these factors?

In Ely *et al.* (2015), they proposed a theory that suspense is in proportion to the expected belief update in the upcoming moment, where the belief refers to the estimated probability regarding a significant consequence (e.g., which team will win the game, which candidate will win the election, etc.). Ely et al. applied this framework to explain the suspense dynamics in different kinds of sports as well as mystery novels, political primaries, auctions etc, but no direct human experiment evidence has corroborated these ideas.

My work sought to develop an empirical paradigm that could test the predictions made by Ely et al. in a controlled but also engaging environment. This paradigm also allowed to compare the "expected future learning" model with other heuristics proposed by the previous literature that I quantified in this setting. In Chapter 1, I present the empirical data as well as the model comparison results in two studies.

1.2 Modeling satisfying explanation as a combination of distinctive causes

In empirical research, scientists constantly face the problem of how to determine which explanation for the data is the best. Statistical methods for model comparison, such as AIC (Akaike, 1974), BIC (Schwarz *et al.*, 1978), Bayesian model selection (Stephan *et al.*, 2009), all aim to balance between the quality of description towards the data (often evaluated in terms of likelihood), and the complexity of the model (could be quantified by the number of parameter, the prior of the model, etc.). This is the "type" level explanation where a myriad of phenomena are summarized and explained by novel theories / models. Plenty of previous psychology studies on causal inference

tasks also explored how people observe or even manipulate instances of causal events, then infer what is the underlying causal structure. Some studies also indicate that probability-based models do well account for people's behavior patterns (Griffiths & Tenenbaum, 2009; Lu *et al.*, 2008).

In contrast to the "type" level explanation, in the daily life, people also often seek for "token" level explanation where the general causal rules are already known, yet they need to find an explanation for a specific instance (e.g. what causes this specific student to be so successful? What disease causes this person to show such a symptom?). How do ordinary people determine the best explanation on the token level? From a computational perspective, do people perform some evaluation strategies similar to the computationally costly statistical algorithms, or do they use some kind of heuristics?

Many previous studies emphasize on heuristic explanation preferences. Specifically, the heuristics regarding preference towards simple or complex explanations have been discussed from different perspectives. People may prefer simple explanations to explain multiple phenomena all at once because they have a bias judging simpler explanations being more probable (Lombrozo, 2007). But if the explanations only probabilistically (not deterministically) causes the phenomena (Johnson *et al.*, 2019), or if the mechanisms behind complex explanation is provided (Zemla *et al.*, 2017, 2020), the explanation preference may shift towards complexity.

In my work, based on the previous research, the aim is to systematically investigate how the different probabilistic settings of the causal system will influence how people prefer a simpler or more complex explanation. By quantitatively manipulate how prevalent and strong each cause is, I can then also compare the Bayesian posterior model of explanation with the behavioral data, as well as developing heuristic models of explanation satisfaction. I will present the novel paradigm as well as analysis in Chapter 2.

1.3 Modeling value-based choice as maximizing a posterior based utility function

When you enter a friend's party, standing in front of tables of snacks, how do you pick which one you will eat first? You may have a vague idea regarding generally how good a general category of snack to you (say, chips are always more attractive than hard candies); Then you may need to more carefully examine a few snacks, comparing between the different flavors, shapes, nutrition contents, etc., to further distinguish which one is better for you. This is a process combining internal preference with external information collection, although in the end still making decisions for one's subjective happiness. What do these two aspects influence the final decisions? How do people integrate their subjective values with sensory information?

Krajbich *et al.* (2010) studied this in a controlled environment and proposed an explanation for the underlying process. They used a paradigm where participants choose between two snack items on the screen, while the experimenters monitoring their eye movement sequences as an approximate of the information collection process. Before the selection phase participants also have seen all the snack images and rated each one, probing the subjective value of each item. It is not surprising that people are more likely to choose the items they rated higher. The surprising finding is that when people look at one item for longer, they are more likely to choose them, on top of the rating difference. Previous studies also have shown that this is potentially causal, i.e. the extended fixation time on items increased the probability of choosing it, not the other way around. How to explain this phenomenon?

The classic treatment from Krajbich et al. (and later extended in Krajbich & Rangel 2011; Krajbich *et al.* 2012 for other types of choice tasks) is the attentional drift-diffusion model where the decision variable is analogous to a drifting particle which goes towards either one of the decision-boundary for the choice options.

In this work, I explore a new, explicitly Bayesian model to explain the same data from Krajbich et al. Specifically, I postulated that when people are looking at one item, they are collecting pieces of information to update the posterior distribution of the item's value. The subjective value rating determines the mean value of each piece of evidence of that given item. Then the posterior distribution is fed into a utility function which includes the posterior mean and variance. The variance term represents people's tendency of being either uncertainty-seeking or uncertainty-averse. Thus the fixation process plays two roles: more evidence collection makes the posterior mean closer to the original subjective value, also makes the variance smaller thus the posterior estimation more certain. In Chapter 3, I will present the details of this novel posterior-based model with technical details of model fitting for choice and fixation data at the same time. I also performed rigorous model comparison with the original and extended version of attentional drift-diffusion model.

CHAPTER 2

Expectations about future learning influence moment-to-moment feelings of suspense

2.1 Abstract

Suspense is a cognitive and affective state that is often experienced in the anticipation of information and contributes to the enjoyment and consumption of entertainment such as movies or sports. Ely *et al.* (2015) proposed a formal definition of suspense which relies upon predictions about future belief updates. In order to empirically evaluate this theory, we designed a task based on the casino card game Blackjack where a variety of suspense dynamics can be experimentally induced. Our behavioral data confirmed the explanatory power of this theory.[1] We further compared this formulation with other heuristic models inspired by studies in other domains such as narratives and found that most heuristic models cannot well account for the specific temporal dynamics of suspense across wide range of game variants. We additionally propose a way to test whether experiencing greater levels of suspense motivates more game-playing. In summary, this work is an initial attempt to link formal models of information and uncertainty with affective cognitive states and motivation.

2.2 Introduction

Suspense refers to sensations of hopeful or anxious anticipation. These familiar affective states often precede the revelation of personally important information—exam results, paternity tests, election outcomes and so forth. However, we also feel suspense in situations where there are no direct personal consequences. For example, children enjoy listening to stories that happen in imagined kingdoms, adults spend time watching televised sports, and Hollywood movies are a

[1]The data that support the findings of this study are openly available in the Open Science Framework (OSF) at https://doi.org/10.17605/OSF.IO/KHPR8.

multi-billion dollar industry. A key feature of these situations is that information is incrementally revealed over time to the observer, often with the goal of building anticipation and arousal.

Relative to the rich palette of our emotional repertoire, suspense is somewhat unique because it is also associated with a strong motivation or desire for information seeking (e.g., finding out what happens, learning the outcome, etc.). Periods of high suspense are known to modulate arousal and attention mechanisms helping to narrow people's focus to relevant stimuli (Bezdek *et al.*, 2015). For this reason, manipulating suspense is a central concern of the multi-billion dollar entertainment industry. Content producers such as movie script writers, video game designers, and novelists all are trained in techniques to increase and sustain the engagement of consumers by strategically manipulating suspense. However, most of these techniques do not derive from a scientific understanding of the nature of suspense as a human reaction to information.

Recently, Ely, Frankel, and Kamenica (2015) proposed a formal (i.e., mathematical) definition of suspense as being derived from the *expectation* that consequential information will be revealed in an upcoming moment. However, their proposed definition of suspense was entirely theoretical. In the present paper we attempt one very specific goal which is to empirically evaluate the merits of the Ely et al. model as a psychological theory of suspense. We design a novel experimental task that enables us to measure people's moment-by-moment perceptions of suspense. By comparing the predicted suspense from the Ely model (and a number of alternatives) to the responses of participants in our experiment we are able to provide a concrete test of theory.

There are a number of unique contributions of this work. The first is that we empirically test a theoretical model about situations that cause suspense, borrowed from the economics literature, using psychological/behavioral methods. The theory itself is a novel approach for the field of emotion because it allows one to make a-priori predictions about how much suspense a person should feel directly from the context of a task or game. We designed a novel paradigm using algorithmically designed stimuli to quantitatively test the theoretical model. Across two experiments we find support for the general principles of the Ely et al. model, though in a slightly different mathematical form. Inspired by previous research on suspense, we also tested several alternative theories which did not provide as good a fit to the data we collected. We conclude with a third experiment assessing if suspenseful games can drive people to play more games even when they receive less or non monetary compensation, connecting the concept of suspense to information consumption behavior.

2.2.1 Previous research

Suspense is a relatively understudied psychological phenomena. However, there are small, but distinct fields that explore the concept of suspense and the relationship between affective states

and information seeking behaviors. The following section reviews some of this research with the goal of situating the unique aspects of the present work.

2.2.1.1 Theories of suspense

Suspense has been studied in many domains of entertainment including narrative literature, film, gaming, and sports. Confirming everyday intuition, the effect of suspense on enhancing entertainment experience has been empirically verified (narrative: Zillmann 1991; sports: Peterson & Raney 2008; Su-lin *et al.* 1997; gaming: Klimmt *et al.* 2009; advertisement: Alwitt 2002).

However, general principles about what drives suspense remains elusive. In the domain of dramatic story-telling, Comisky and Bryant (1982) propose that suspense will be higher if 1) the audience disposition toward the protagonist is more positive and 2) the belief that the protagonist will fail is higher.

A related concept is the narrative device of removing a possible solution to a problem facing the protagonist (Gerrig & Bernardo, 1994) and situations where a negative event becomes more likely (e.g., in a game, Klimmt *et al.* 2009). Suspense is also known to increase when the audience knows something is about to happen while the character does not (Alwitt, 2002). The temporal dynamics of the narrative also matters. Alwitt et al (2002) propose that the presence of time pressure and alternations between moments of hope and fear before the resolution also increases suspense. The principles identified in these reports are undoubtedly powerful moderators of suspense. However, the deeper principles remain largely qualitative and domain specific.

There have, however, been some attempts to unify the definition of suspense across a broader set of situations or domains. For example, Lehne & Koelsch (2015) attempts to unify the suspense in narrative stories and in music (under the concept "tension"). They developed a domain-independent model stating that suspense "originate(s) from states of conflict, instability, dissonance, or uncertainty that trigger predictive processes directed at future events of emotional significance". The more divergence there is between possible future outcomes, the more suspense is generated. Although a more concrete definition of this divergence is lacking, this theory does highlight some critical psychological components of suspense, particularly the notion of uncertainty and predictive process. en there are diverging possible outcomes for the immediate moment.

2.2.1.2 Information anticipation and uncertainty

While the narrative devices used to build up anticipation in a story are complex and involve many aspects of semantic knowledge, the widely used paradigm of conditioning could be seen as a much simpler form of manipulating an organism's anticipation. Classical conditioning involves the learned anticipation of a positive or negative outcome following an unconditioned stimulus (US),

such as an audio tone which occurs repeatedly before an electric shock (Pavlov, 2010). The period of waiting for the stimulus to arrive may, at least intuitively, involve some of the same emotional feelings of suspense including anxiety and a strong desire to have the uncertainty resolved.

In addition to its simplicity, one advantage of such paradigms is that they allow more careful measurement of how information seeking is modulated by uncertainty and anticipation. Unlike in narrative settings where the uncertainty is hard to quantify, in the conditioning paradigm, researchers are able to manipulate the relationship between cues and rewards thus introducing different levels of uncertainty (White & Monosov, 2016). Monkeys are attracted to visual cues that will resolve uncertainty about future rewards. For example, they are more likely to shift their gaze to informative cues rather than cues signaling more rewards but with no uncertainty (White *et al.*, 2019), indicating the importance of uncertainty reduction for animals. The neural networks responsible for the expected uncertainty resolution have begun to be identified (White *et al.*, 2019; Horan *et al.*, 2019).

Uncertainty-driven arousal and anticipation is not limited to personally experienced outcomes. We experience suspense in movies and stories somewhat vicariously: the fate of the protagonist for instance is not our own. Relatedly, there are a number of experiments exploring behavioral and neural responses to vicariously experienced rewards and punishments (known as social conditioning). For example, if a subject in a conditioning experiment simply observes a video of another subject who they believe is experiencing painful shock, they begin to experience similar levels of anticipatory arousal (Olsson & Phelps, 2007). Similarity and relatedness between the viewer and the foil will enhance the strength of vicarious learning and arousal, e.g., in humans an in-group bias is also present (Golkar *et al.* 2015; see Debiec & Olsson 2017 for a comprehensive review). These findings draw some parallel to the fact that building empathy towards a protagonist is a useful tool for invoking greater suspense.

Besides the similar information delivery structure between conditioning paradigms and suspense-inducing scenarios, the elicited emotion is also closely related. Conditioning is a critical tool for researchers investigating emotions like fear and anxiety in humans (Maren, 2001), that are also closely related with the feeling of suspense (Nomikos *et al.*, 1968). However, as described below, the theory of suspense we develop applies to much more complex situations than the uncertainty about the timing or delivery of a reward or punishment and captures rich temporal dynamics of suspense over time.

2.2.1.3 Non-instrumental information-seeking, or curiosity

Besides being an emotional state, suspense often acts as a motivating force for active information-seeking behavior. Examples already considered include how movies and other media attempt to hold our attention by increasing suspense at key moments.

Conceptually, factors that promote intrinsically motivated information-seeking - sometimes called curiosity (Gottlieb & Oudeyer, 2018; Loewenstein, 1994) also relate to suspense. In line with Berlyne's categorization (Berlyne, 1966), suspense is more linked to "specific exploration" a.k.a information-seeking towards a specific object (as opposed to "diverse exploration", more driven by novelty-seeking, surprisingness, complexity and so on). Suspense is usually about specific questions, like "who will win the game?" or "will the protagonist get killed?" Many recent theories of curiosity reinforce the parallel with suspense. For example, Lowenstein (1994) claims that curiosity is a result of an "information gap." For suspense, the gap naturally exists when the audience cares intensely about the result of something but the information is still not provided yet. Quantitatively, van Lieshout et al. (2018) found that more uncertainty also increases the level of self-reported curiosity as well as eagerness to view an unrevealed outcome.

Despite this work, specific evidence for the impact of suspense on information-seeking is lacking (but see Bezdek *et al.*, 2015). What we do know is that suspense makes sports games (Peterson & Raney, 2008; Su-lin *et al.*, 1997), stories (Zillmann, 1991), and commercial advertisements (Alwitt, 2002) more enjoyable and enjoyment could be a mediating factor for the further information-seeking or consumption. In our study, we will test the behavioral effect of suspense in a more controlled setting, aiming to assess the degree to which experimentally induced feelings of suspense influence the desire to further information consumption.

2.2.2 Theory: Suspense as the expected future belief update

A recent paper proposes that suspense can be defined as an increasing function of the "expected future belief update" (Ely *et al.*, 2015). Here the beliefs refer to the subjective probability of a significant outcome (e.g., which team will win a game) that is updated over time with information as an experience unfolds. For example, while watching a game we might form the impression that there is a 60% chance our favored team will win given the current score and time clock. In addition to tracking their momentary belief, people are assumed to also estimate how their belief may change in the future (a "prospective" type of calculation). For example, if a doctor arranges to call a patient at a particular time with test results, in the period leading up to the phone call the patient might expect that their belief about their health could soon change (although they may not know what they will learn). Conditioned on the information one expects to receive, if the subsequent future beliefs would be very different from one another they would be said to have high variance. For example, if the test the doctor performed was routine, the patient would not expect their future knowledge state to change much after the call (low variance). As a result they would experience low levels of suspense. In contrast, if the test was a cancer screening, then the call might either alter the person's life or leave them reassured (high variance), and thus they would

experience high levels of suspense in that moment.

To formalize these intuitions, following Ely et al., we assume that a viewer's subjective belief μ evolves over a series of discrete time points t, such as individual points in tennis, card draws in a game, or (discretized) time passing in a movie. At each time point, relevant information may be encountered and people update their belief μ_t (e.g., by Bayesian updating). In addition, viewers also anticipate future information using their understanding of the situation. For example, a viewer might anticipate that their favorite team will score on the next play or that the opposing team will score, each representing a state s. The state s has a probability of being realized $P(s)$ (determined by things like the mechanics of the game and the abilities of players) and will result in a future belief μ^s_{t+1}. The variance among these beliefs indicates how different the future might be, and therefore how much suspense might be evoked.

Formally, Ely et al. defined the momentary suspense at time t, $\mathcal{S}_t$ as:

$$\begin{aligned} \mathcal{S}_t &= \mathrm{E}_s[(\mu^s_{t+1} - \mathrm{E}_s[\mu^s_{t+1}])^2] \\ &= \mathrm{E}_s[(\mu^s_{t+1} - \mu_t)^2] \\ &= \sum_s P(s)(\mu^s_{t+1} - \mu_t)^2 \end{aligned} \tag{2.1}$$

The term $\mathrm{E}_s[\cdot]$ represents the expected value of a quantity averaged over all values of s. The step from line 1 to 2 of the equation is because $\mathrm{E}_s[\mu^s_{t+1}] = \mu_t$ (i.e., the belief now is the expectation over all possible future beliefs; derivation in Appendix material).

Note that the difference $\mu^s_{t+1} - \mu_t$ indexes a quantity we might associate with surprise in the sense that it is the difference between what one thinks now compared to after learning a new piece of information.

As a result, the value $\mathcal{S}_t$ can be also be interpreted as the expected future surprise or expected future belief change from the current to the next time period. The phrase "expected surprise determines suspense" is a good summary of the intuitive implications of the theory.

Figure 2.1 gives a graphical overview of the model applied to a hypothetical tennis match. Here μ is the probability of winning the match ($\mu = 1$ if player A is certain to win and $\mu = 0$ if they are certain to lose), each point is one time step, and s is whoever wins the next point. In the center of Figure 2.1 we show the unfolding of belief about who will win for two different matches (1 and 2) with the x-axis representing time. The panel on the left shows why the beginning of both matches is not very suspenseful: whoever wins the first few points has little impact on predictions about the final outcome given how much time remains in the match. However, the end of match 1 is predicted to be more suspenseful since whoever wins a point will greatly swing the final outcome (indicated by the top right panel where μ_{t+1} is quite different, or variable, depending on what happens), while

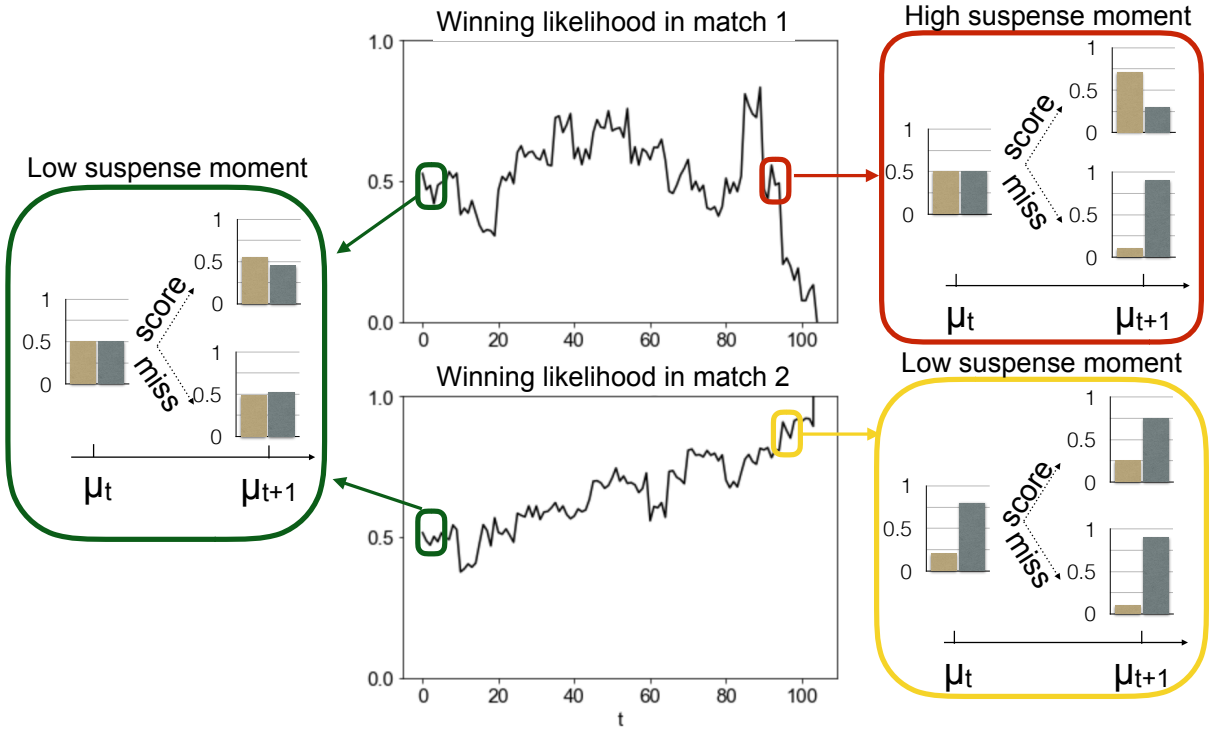


Figure 2.1: The evolution of belief (i.e., likelihood of player A winning) over two different tennis matches. Different moments during the match may correspond to different levels of suspense, that we will explain in terms of a difference in the expected future belief update (colored boxes). **Left (green box)**: At the start, suspense is low because the potential updated beliefs for the next time point do not differ dramatically; **Right top (red box)**: Here is a moment of high predicted suspense in which the next point is expected to have a major impact one way or another on the outcome; **Right bottom (yellow box)**: here player B is already very likely to win, making the expected belief update small thus resulting in low suspense.

match 2 is less suspenseful since one side has already virtually secured victory and so no matter how the next point plays out overall expectations about the match remain the same (bottom right panel, μ_{t+1} is similar no matter what happens). The fact that in match 1 there is more suspense at the end of the match rather than the beginning (despite the overall belief about who will win being somewhat uncertain) shows an important feature of the theory. It is not simply how uncertain you are at the current moment, it is how much you expect the information in the next moment to change your belief.

In the following studies, we will introduce different implementations of this theory, then compare it with other suspense models inspired by the previous research. Details are provided in Experiment 1.

2.3 An experimental test of the theory

Ely *et al.* (2015) articulated the basic outline of the theory described above and explored a number theoretical analyses of the optimal structure for games to maintain suspense. However, to our knowledge, this definition of suspense has not yet been examined in psychological experiments (although Wilmot & Keller 2020 conducted an examination using natural language processing techniques). We propose that a useful behavioral paradigm for testing this theory needs to have at least two features:

1. The experiment context should be quantifiable in a probabilistic model. This tends to exclude tasks like reading stories and watching movies because it is not trivial to convert these complex situations into accurate probability models. This limitation is more practical than conceptual however.

2. The experiment paradigm should allow the decoupling of the external stimulus and internal belief. In most prior work, changes in suspense are always confounded with incidental features of the stimuli. For example, suspenseful moments of a movie might have more visual motion. To validate the belief-based account of suspense, the ideal experiment would manipulate an internal belief through some prior knowledge while holding other aspects of the stimulus and task identical.

With these criteria in mind, we designed a single-player card game related to the classic casino game Blackjack. Participants randomly draw cards from a small deck with a known distribution of cards and report their moment-by-moment suspense. If the sum of cards exceeds or hits a critical threshold the game is lost. Unlike in the Blackjack, the participant does not make decision on when to stop drawing cards but instead is limited by the maximum card number depending on the rules. Thus the participants win by chance. Intuitively, suspense builds in the task when the sum of the drawn cards approaches the critical threshold (e.g., when the sum of drawn card exceeds 21 in Blackjack). Because the distribution of cards and the probability of drawing any card can be determined exactly, the game is an ideal test bed for exploring information-theoretic models of suspense, including the Ely theory. In addition, the game is relatively fun, intuitive, and easy to explain to participants.

To address the second concern from above, in Experiment 2, participants were given one of two different rules for how the game would be scored, thus matching the stimuli while changing their implications for the outcome. In one version, the game ended in a loss if, at any point, the sum of the cards drawn so far met or exceeded the threshold. This is the traditional concept of "bust" from Blackjack. In a second version, the game ended in a loss only if the sum met or exceeded the boundary value on the final draw of the game. Since we allowed the presence of negatively

valued cards, it was possible, under the second rule set, for the sum to exceed and then return to safety below the threshold before the game ended. The differences between these two rules allow us to compare identical sequences of cards, but to modulate if a given card draw is more or less suspenseful about the game outcome according the the Ely et al. theory. To optimize the power of our experimental approach, we used a computer-aided search to find a combination of rules, decks and card sequences that resulted in strong predicted suspense differences between the two rule sets.

2.3.1 Overview of the experiments

In total, we present 3 experiments using this paradigm. Experiments 1 and 2 test how well the suspense model predicts participants' moment-by-moment subjective report of suspense. Experiment 1 establishes the theory's explanatory power across a relatively large variety of stimuli. At the broadest level, we contrasted games that the model predicts will be extremely low in suspense with games predicted to produce high overall suspense. If the model is even reasonably in line with human suspense, these differences should appear robustly. Second, the model makes point-by-point predictions about the fluctuations in suspense within a game. We used model comparison to assess how well the Ely et al. model accounts for these fluctuations compared against a number of variants heuristics and baselines. Experiment 2 is designed to introduce different temporal dynamics of rules to test the theories in broader contexts. Experiment 3 then asks whether people's willingness to play more games is affected by the level of suspense they experienced in earlier games.

2.4 Experiment 1: Predicting the dynamics of suspense

The goal of Experiment 1 is to provide an initial evaluation of the Ely et al. model. We constructed a wide variety of games and compared the moment-by-moment subjective ratings of suspense from participants with the predicted levels of suspense from the model (and related model variants).

2.4.1 Experimental Method

2.4.1.1 Participants

We recruited 191 participants (age M = 36.2, SD = 13.2, 96 female, 5 undisclosed gender) from Amazon Mechanical Turk using psiTurk (Gureckis *et al.*, 2016). Participants were offered a $0.30 base payment plus the option of a bonus (all participants ended up receiving a bonus between USD

$0.60 and $1.20 as described below). Half of them (96) were randomly assigned to the "high predicted-suspense" condition (described below).

We decided the participant number with the expectation of at least 15 participants per point.

2.4.1.2 Procedure and game design

The instructions and main task were completed by participants on their personal computers using a custom javascript interface in the browser. The task took around 14 minutes (SD=3).

Participants were told that we were interested in their feelings of suspense while playing a simple card game. Each participant went through an extensive tutorial covering the rules of the game and could only continue if they correctly answered a series of comprehension questions to make sure they understood the rules. They then played a training game that was identical to the test games except there was no bonus attached to a win. After completing this, participants played three games. A $0.60 bonus payment was earned for each game that the participant won. As described below, all participants won either one or two games because the outcomes were, in reality, fixed and not under their control, although the task was designed to make it seem to participants that were playing a game of chance.

Similar to Blackjack, in each round of a game, the player drew cards from a deck. In this case, decks contained nine cards with visible values (Figure 2.2A). To increase the trial-by-trial (i.e. point-by-point) suspense dynamics, we used the following two stage process for revealing each card: First, the participant saw the face value of the nine cards in the deck. Then, the cards were flipped over and an animation was shown of the cards being spatially shuffled (Figure 2.2B). Next, two cards at the top of the deck following the shuffle were selected and moved to the left hand side (Figure 2.2C). At this point on a randomly selected subset of trials (around 60%), a self report of suspense was elicited (Figure 2.2D). Next, the participant pressed a button on the keyboard to spin an animated wheel that determined the actual identity of the final card to be drawn (Figure 2.2E). Participants could choose how long to spin the wheel (by holding down a button), but in actuality the spinner always landed on the card determined by our chosen game sequences. The purpose of the spinner was to give participants an (illusory) feeling that the outcome was truly stochastic and that they could potentially use skill of spinner control to obtain a more favorable outcome (increasing engagement with the task). After a card was selected, the participant's current card total (the sum of the face value of all of the cards they had drawn so far) was automatically updated in a graph at the top of the screen (Figure 2.2F). The interface displayed the total sum as well as the graphical history of the sum as it evolved across the sequential draws.

To measure suspense, after the two candidate cards are shown but before the process of spinning the wheel, we asked the participant to rate their current suspense using the keyboard numbers 1 to 5, where 1 means "no suspense" and 5 means "high suspense". Previous studies on subjective

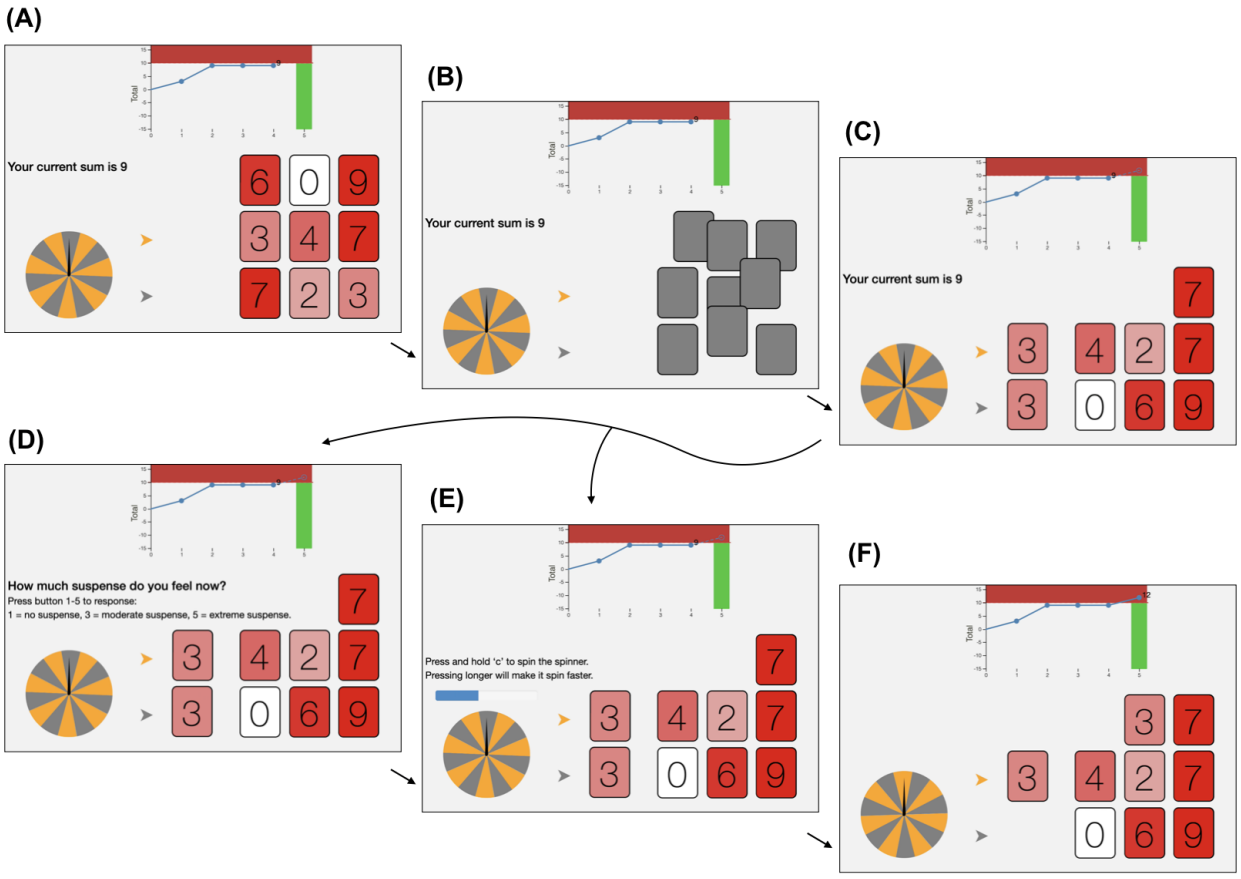


Figure 2.2: The game interface and animation sequence for a single card draw. (A) The deck and current sum are revealed (B) The cards are flipped over and shuffled (with animation) (C) The first 2 cards after shuffling become "focal" candidates (D) Occasionally, self report of suspense are requested from the participant (E) Participants press a button to "control" the spinner speed (F) The final position of the spinner determines which card is finally chosen.

reporting of suspense have also used 7-point (Gerrig & Bernardo, 1994; Knobloch-Westerwick *et al.*, 2009) or 11-point (Cupchtk *et al.*, 1998; Comisky & Bryant, 1982) scales, yet we are unaware of any psychometric comparison of different response scales for suspense measurement. No other instructions were given about the use of the scale. However, we asked participants to describe how they personally understood the term "suspense" in the post-task questionnaire. To minimise interruption of the game experience, suspense rating were only requested on 2 or 3 randomly selected draws per game.

At the end of each game, participants were given the option to play one more game with only half the bonus of the previous games. This was not required and participants were free to decline to play. This final game was intended to test whether the overall amount of suspense encountered in previous games affects people's willingness to play additional games. More comprehensive

explanation and analysis of this data set, along with other contrasting conditions, are presented with Experiment 3.

Finally, participants were presented a questionnaire asking: 1.) If they found practice round helpful for their understanding (range of 1-4), 2.) Whether they felt the game was fair (range of 1-4), 3.) How they judged suspense (free response), 4.) How they decided whether to play the additional game (free response), 5.) If they found any problems or had any suggestions about the whole task (free response), and 6.) Demographic information (gender, age, education). Most of these variables were not analyzed because we had no specific hypothesis (e.g., about gender or age) but they are reported here to facilitate secondary open science data analysis. Participants reported voluntarily and it was still possible to submit the task to Mechanical Turk if the questions were left blank.

2.4.1.3 Stimuli selection

In this task, the potential stimuli space is very large. For a given deck of nine cards, there are 60,466,176 sequences of five draws of two ordered cards with replacement, not to mention practically unbounded variation of deck compositions. This raises the issue of selecting a set of stimuli that will best serve to test the theory.

We searched the space of games by generating 5000 random combinations of deck and card draws that were valid under the game rules, before filtering with several heuristic restrictions to ensure the games also felt like a plausible random draws from the deck. For example, one heuristic restriction prevented drawing the same card more than three times in game. We then calculated the suspense for each draw of each game based on the suspense model (described below) before further filtering the games to find games with either high or low predicted suspense, and both winning and losing outcomes. The full selection procedure is detailed in the Appendix materials.

As mentioned, participants were randomly assigned to either a "model predicted" high or low suspense condition. In the high predicted-suspense condition, all games were selected from the 2% highest suspense games among all the simulated games; in the low predicted-suspense condition, all games were selected from the 10% lowest suspense games. The difference in thresholds for inclusion between the two conditions is because for games end up losing they usually have relatively high level of suspense.

Furthermore, we selected for the outcome of games so that participants experienced either one or two wins out of the three games they played. This ensured that the influence of game outcome on suspense was relatively neutral across participants.

2.4.2 Modeling Approach

2.4.2.1 Belief updating process in the card game

The Ely et al. theory does not explicitly specify the belief updating model that would apply to this game. However, given the simplicity of the game dynamics it is possible to create a Bayesian updating model that is exact. In particular, to calculate the belief μ_t (probability of winning at a given moment), we used an exact enumeration approach, counting all the possible future card draws and their respective win or loss outcomes, which generated an overall win probability. For example, in a game with five total cards draws, if the sum after the fourth card draw is 7 then one can simulate all possible subsequent draws and tally how many result in a win (e.g., card sum $<$10) or loss (e.g., card sum $\geq$10). This determines the probability of going on to win the game given what is known at time point t.

Since the suspense is reported after the pair of potential next cards are drawn, we can calculate the probability of winning (μ_{t+1}) once one of these two cards cards is selected (using the same process as for calculating the current belief). The suspense prediction for this time point then follows equation 2.1. Given the wheel has equal area for both options, the probabilities of both future states are equal (i.e., $p(s) = 0.5$) and so suspense here simplifies to the variance of μ^s_{t+1} over the two possible outcomes s.

2.4.2.2 Model-based data analysis

We introduced a response probability model to convert the continuous suspense predictions from the model (on the range 0.0 to 1.0) to a integer output in the range of 1 to 5. This allowed us to estimate the likelihood of a given subject's response R_t on a given trial t given the predicted suspense $\mathcal{S}_t$ (i.e. $p(R_t|\mathcal{S}_t)$). The response model has a single noise parameter optimized to maximize the likelihood of producing the behavioral data. See the Appendix materials for the full details.

For each individual we collect only 8-9 data points, which is too few for individual modeling, thus we analyzed the data at the group level by using the group parameter for likelihood fitting, and averaging the response given on each point for the correlation analysis. Since each participant was randomly assigned to different games and asked to indicate their suspense at randomly selected points in the game, each averaged data point is based on a slightly different number of participant responses (M = 40.5, SD = 4.0).

2.4.3 Alternative models

Before describing the results of Experiment 1, we consider two types of alternative models: 1) Variations of Ely et al. model that also rely on expected future belief calculation but measure the

probability change in ways other than squared distance (as in Eq 2.1). 2) Heuristic models inspired by previous qualitative research on suspense (introduced in "Previous research" section).

2.4.3.1 Different measures of suspense

To measure suspense as expected belief change, Ely et al. used the squared distance between probabilities before and after encountering some new information. Yet the justification for this choice is somewhat unclear. There are many ways to calculate the how far or how much a belief has changed and so we consider a number of other alternatives (see a Nelson et al. 2005 and 2010 for an extensive discussion of these issues).

To recap, in the Ely et al. model suspense is defined as the expected squared distance, which is also known as a L2-norm, for the belief update on the next time point:

$$S_{\mathrm{L2}} = \mathrm{E}_s[(\mu^s_{t+1} - \mu_t)^2] \tag{2.2}$$

where $s = 1, 2$ for each possible card to be drawn and $\mathrm{E}[\cdot]$ denotes the average over s.

We additionally explore alternative metrics to quantify the belief updates. For example, entropy reduction is defined as follows:

$$S_{\mathrm{H}} = \mathrm{E}_s[H(\mu^s_{t+1}) - H(\mu_t)] \tag{2.3}$$

where H denotes the Shannon entropy:

$$H(p) = p\log(p) + (1-p)\log(1-p) \tag{2.4}$$

Alternatively, we could use an absolute error norm, or L1 norm:

$$S_{\mathrm{L1}} = \mathrm{E}_s[\mathrm{abs}(\mu^s_{t+1} - \mu_t)] \tag{2.5}$$

The absolute error norm is closely related to concepts of "probability gain" and "impact" Nelson *et al.* (2010).

A last form is similar to the L-1 norm but a more statistically justified form called the Hellinger distance which is a special case of f-divergence:

$$S_{\mathrm{Hellinger}} = \mathrm{E}_s\left[\mathrm{Hell}(\mu^s_{t+1}, \mu_t)\right] = \mathrm{E}_s\left[\frac{1}{\sqrt{2}}\sqrt{\left(\sqrt{\mu^s_{t+1}} - \sqrt{\mu_t}\right)^2}\right] \tag{2.6}$$

Previous studies provide evidence suggesting human prospective informativeness judgments may be driven by an absolute error norm (Nelson *et al.*, 2010) while there no clear difference between

the other information-theoretic norms (Nelson, 2005).

Intuitive demonstrations of how these mathematical forms differ are provided in Appendix materials (see Appendix Figure 2.13 and 2.14).

Note these different variations are all under the conceptual framework that suspense is based on expected future belief update at the next time point. They are only different in how the magnitude of the belief update is measured.

2.4.3.2 Heuristic models

We now introduce two conceptually different models of suspense inspired by the literature: Uncertainty and "Fear of losing".

Many researchers agree that higher suspense is related to high uncertainty (E.g. Lehne & Koelsch 2015; Although note the "paradox of suspense" Yanal 1996 where uncertainty is not always necessary). Also in the realm of psychology, uncertainty has been found to sustain attention since people may desire the reduction of uncertainty on a motivational level (Berlyne, 1960).

Quantitatively, we define this model in two ways: the entropy of current belief distribution (Uncertainty, probability based) and card difference (cardDiff, pure heuristic), detailed in the appendix.

The last alternative theory we consider is that people may feel more suspense if the negative outcome is very likely to happen. Previous studies in film narratives (Comisky & Bryant, 1982) and sports viewing (Knobloch-Westerwick *et al.*, 2009) both empirically found that when there is a greater chance for the unwanted outcome to happen, more suspense is experienced.

Mathematically, we introduce two models where suspense is higher when probability of winning is lower. These models are identical except that they deal differently with the cases where the probability of losing is certain. Specifically, when probability of winning is zero, it could either be maximum suspense (pLose model) or zero suspense (almostLose model). Thus the almostLose model includes a discontinuity such that the suspense increases as probability of losing increases but falls to zero when losing is certain. The exact formulation and other heuristics related to this idea are detailed in the Appendix "Alternative models" section.

To sum up, we formulated eight models that derive from three different intuitions about suspense as summarized in the Table 2.1. We compare these models to see which best accounts for the behavioral data.

2.4.4 Results

In Experiment 1, we first examined the effect of the high versus low predicted suspense manipulation. This provides a sanity check whether our stimuli lead to clear differences in self-reported

Table 2.1: Alternative Models

Principle	Implementation	Name
Future belief update	Squared belief difference	L2 (Ely)
	Absolute belief difference	L1
	Entropy reduction	ΔH
	Hellinger distance	Hellinger
Uncertainty	Entropy of current belief	uncertainty
	Card difference	cardDiff
Fear of losing	Probability of losing	pLose
	pLose with ceiling	almostLose

suspense. Then, we quantitatively compared the Ely et al. model and other alternatives (all models listed in the Table 2.1) to the behavioral data.

2.4.4.1 High/Low predicted-suspense manipulation

We first checked whether the participants in the low predicted-suspense condition indeed reported lower suspense on average than those in the high predicted-suspense condition. The distributions of suspense responses are clearly separated in two conditions (shown in Figure 2.3). One-sided Mann-Whitney rank test confirms that suspense reported in the high-predicted condition is significantly higher (U = 3.0e5, $p < 0.001$). Participants in the low predicted-suspense condition reported the lowest level of suspense most often, whereas in the high predicted-suspense condition the highest level of suspense was most often selected. Thus, our paradigm had successfully manipulate the feeling of suspense by designing specific decks and card sequences.

Note that in Figure 2.3 the discontinuity in the suspense distribution may due to the instruction about keyboard responding being a little misleading (see Figure 2.2D) which we recommend should be avoided by future researchers).

2.4.4.2 Suspense dynamics

To visualize the detailed game suspense trajectories, we plotted the average suspense reports at each time point in Figure 2.4 (high-predicted suspense games) and Figure 2.5 (low predicted suspense games). To compare these against the predictions of the model variants in Table 2.1, we also plotted the suspense from each class of model: future belief change (L1), uncertainty and fear of losing (almostLose). Note these predictions are parameter-free thus not fit to the data in any way.

From these trajectories we can observe several features. First, there is considerable agreement between participants' suspense judgement such that the aggregate suspense trajectories are not flat

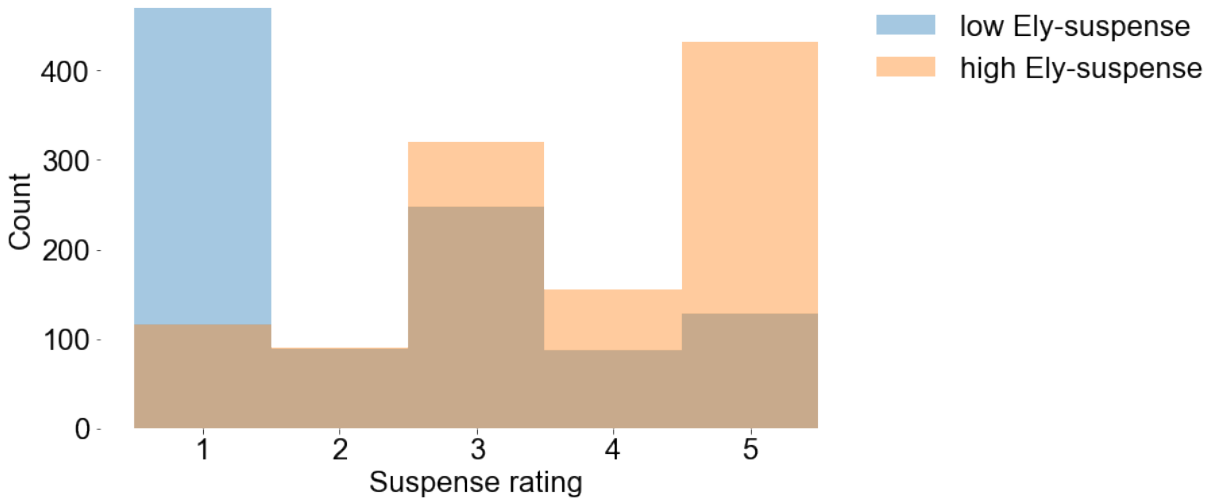


Figure 2.3: Experiment 1, Distribution of reported suspense in low and high predicted-suspense conditions.

Table 2.2: Experiment 1: Model Comparison.

	Future belief update				Uncertainty		Fear of losing	
	L1	Hellinger	L2	ΔH	uncertainty	cardDiff	almostLose	pLose
Pearson's r	0.82	0.82	0.74	0.74	0.74	0.68	0.65	0.54
Likelihood difference	7.47	6.43	3.97	3.85	8.88	3.55	4.56	2.71

Note: 1st row: Correlation coefficient between the parameter-free model prediction and the raw participant report of suspense.
2nd row: Maximum Likelihood Estimate fitting of all individual trials above the baseline (random report), averaged to per trial, in the unit of 10^{-2}.

but shows clear within- as well as across-game dynamics. Second, the L1 model shows qualitatively good tracking of the trends in the high predicted-suspense condition (Figure 2.4). However, in the low predicted-suspense condition, the L1 and uncertainty models systematically underestimate participants' reported suspense level while almostLose heuristic is significantly modulated by if the game leads to a win or loss (Figure 2.5).

To quantitatively test how well the models explain the dynamics of suspense on a point-by-point level, we first checked point-by-point correlation between averaged participant suspense judgments and the parameter-free model predictions across all the games. The correlation coefficients for all models are listed in Table 2.2. The models with the highest correlation are the two "future belief update" models Hellinger and L1 heuristic model, followed by other models in the same family (L2 and ΔH) and the Uncertainty heuristic. Interestingly, all models from the "future belief update" family give higher Pearson r than heuristic models. Figure 2.6 shows the joint distribution

of reported suspense ratings per trial (averaged across people) for examples from each of the three classes of model types.

We also used maximum likelihood estimation to fit a probabilistic version of each model using a single "response noise" parameter to all responses points (N = 2139). Details for how we mapped the model predictions to response scale elements are given in Appendix section ''Computing model likelihood"). The per-trial likelihood improvement for the given model relative a null model which assumes a uniform probability of each response category is presented in Table 2.2 (bottom row). According to this measure, the uncertainty heuristic fits the best, followed by the L1 and Hellinger model. The "fear of losing" models perform relatively worse in both correlation and likelihood. The likelihood fits differs from correlation in requiring a match not just in terms of covariance but also in terms of absolute values of the ratings being calibrated correctly. Also the likelihood fits all the individual trials whereas Pearson'r is against the group average.

2.4.5 Discussion

Our new experimental paradigm allowed us to elicit reasonably consistent suspense reports between participants. The overall model-based high/low suspense manipulation had a clear effect despite the inherent noise in self-reported data.

We then tested several classes of suspense models with specific variants of each. Overall, we found that the "future belief update" models, especially those using the L1 and Hellinger metrics, described the data best. This supports the broad idea at the heart of the Ely et al. proposal that suspense indexes anticipation of future belief changes. However our results are suggestive that the specific formulation of this anticipation (using the L2 norm) may not best characterise human suspense.

Among the (non-expectation-based) heuristics, the current uncertainty model provided the best likelihood fit. It is surprising that the simple Uncertainty model fits the data well since it is insensitive to the whether and when the uncertainty can be resolved. For example, this model assumes that the very beginning of a game can induce th same level of suspense as much as the match point, as long as people are similarly uncertain about the outcome. This seems counter-intuitive and we wonder whether this will generalize beyond the current result since experiment 1 only explored a limited number of games and, more importantly, under a specific game rule.

A key prediction of the Ely et al. approach is that suspense is a function of expected belief rather than something evoked by any particular stimulus or situation. One way to test this hypothesis is to show people games which are identical (in terms of the face value of the cards being drawn) but to manipulate, across conditions, their belief about whether they are close to winning or losing. To this end, in Experiment 2, we designed two games for each of the same deck and card sequence

such that the difference is only in the "rules" of the game that determine when the player wins. This manipulation enables studying the mechanism driving suspense without the potential confound of specific card and deck information. In addition, this provides a replication and a new test of the model's generalizability.

2.5 Experiment 2: Rules with different temporal structures

To disentangle the alternative theories of suspense and examine their performance in a wider variety of game structures, we introduced a pair of alternative game rules:

1. The "Bust" rule. This is same as in Experiment 1 where the game is lost any time the sum of the cards drawn so far meets or exceeds the boundary value.

2. The "No Bust" rule. Here the game is lost only if the sum meets or exceeds the boundary value *on the final draw of the game*. Due to the presence of negatively valued cards, it is possible in this game for the sum to exceed the bound but then return to safety by the final draw.

Using the same card game paradigm except adjusting the rules, we could then contrast the suspense response when playing games with exact same card sequences but under different rules, highlighting how suspense is modulated by the internal belief.

2.5.1 Method

2.5.1.1 Participants

We recruited 263 participants (age M = 36.7, SD = 20.4, 113 female) from Amazon Mechanical Turk using psiTurk (Gureckis *et al.*, 2016). They were paid \$0.90 (\$.30 base pay and a \$.60 bonus which was, in fact, the same for all participants). The task took 12$\pm$3 minutes to complete. There were 144 participants assigned to the "Bust" rule and the rest to "No Bust" rule condition. We decided the participant number with the expectation of at least 80 participants per point. This number is bigger than experiment 1 because we expect the effect of rule manipulation will be much smaller than the high / low suspense game manipulation.

2.5.1.2 Procedure and stimuli

The games and interface were as in Experiment 1 except that there were just 3 draws per game rather than 5. Each participant was assigned to one of the two rule conditions and played 2 practice games then 3 actual gambling games.

2.5.1.3 Rule design and model-based stimulus selection

To optimise the power of the design to distinguish between candidate models, we used computer-aided search to find game sequences (i.e., card decks and the sequence of card draws) which are valid under both rules but also lead to robust differences in predict suspense according to the Ely et al. type belief models.

To implement this design, we searched for a diverse set of games with large predicted suspense differences by maximising:

$$\begin{aligned} \text{score}(\text{seq}, \text{deck}, \text{rulepair}) = \mathcal{S}^{\text{rule1}} + \mathcal{S}^{\text{rule2}} \\ -\alpha \cdot r(\mathcal{S}^{\text{rule1}}, \mathcal{S}^{\text{rule2}}) \end{aligned} \tag{2.7}$$

where α is a positive weight constant and $r(\cdot)$ is Pearson's correlation coefficient. The first two terms ensure the average suspense level is not too low while the third encourages anti-correlation between the suspense trajectory under two rules given the Ely et al. model. We recomment setting α to a positive constant around 1 which ensures the two terms have similar magnitude.

We then searched the space of games to select valid stimuli as we did in Experiment 1. For this rule search, we added some additional steps. We searched the rule space for pairs of bust and no-bust rules with different boundary values, then generated random game combinations. We then selected the top 10 games of each rule pair, calculated the average score of these games then picked the rule pair with the best scores. The final winning rules were: bust with a bound of 7 (i.e. the card sum should never exceed 7) and no-bust game with a bound of 3 (i.e. the sum of cards should not exceed 3 at the end of three draws).

Each participant was assigned to one rule condition and played two training games (meaning they did not contribute to the participants' final bonus) then three test games with a potential \$0.30 bonus. Of these, two were overall high predicted-suspense and one was low predicted-suspense. The order of games were all counterbalanced.

2.5.2 Results

The goal of Experiment 2 was to isolate prospective beliefs from other features of our stimuli by matching game sequences to be identical but changing their implications with our rule manipulation. This allows us to better study what drives feelings of suspense even if the game sequences were the same, and how well the different model categories capture that effect. Specifically, we wanted to know if the Uncertainty heuristic model which is not sensitive to the rule's temporal constraints will describe the behavioral data as well.

We first checked the correlation between the game suspense trajectories (averaged over all

Table 2.3: Experiment 2: Model Comparison

	Future belief update				Uncertainty		Fear of losing	
	L1	Hellinger	ΔH	L2	uncertainty	cardDiff	almostLose	pLose
Pearson's r	0.91	0.90	0.86	0.86	0.83	0.77	0.61	0.26
Likelihood difference	11.07	10.64	8.90	8.96	5.96	6.32	2.28	-0.00

Note: 1st row: Correlation coefficient between the parameter-free model prediction and the raw participant report of suspense. 2nd row: Likelihood fitting all individual trials minus the baseline (random report), averaged to per trial, in the unit of 10^{-2}. All the heuristics are worse than the "future belief update" models. The best model is the "future belief update' models of L1 and Hellinger form.

participant response) and the parameter-free model predictions. It turns out that all "future belief update" models have higher correlation as well as likelihood than the other heuristics models (both listed in Table 2.3; Correlation see Figure 2.7). Among the "future belief update" models, metrics forms of L1 and Hellinger again perform the best as in Experiment 1.

In addition, we assessed how well the models predict the suspense difference between the two rule conditions for identical card sequences. This analysis compares the difference in the model prediction for each game point under the two rules against the difference in empirical suspense ratings, with the empirical suspense aggregated among all the participants in the same rule condition. Note the low predicted-suspense games are excluded in this analysis since the suspense in different rules are supposed to be the same.

Figure 2.8 shows the suspense under 2 rules in each game plus their difference. From Figure 2.8C, we can see that although the card sequences and decks are identical, the suspense reported by participants clearly depends on the rule. In several places the difference significantly departs from zero i.e. the 95% confidence interval does not include 0. In those cases, the L1 and Uncertainty models predict this rule difference in the right direction, while almostLose model does not.

Figure 2.9 visualises the match between model predictions and judgments. The L1 and Uncertainty models predict the direction of rule-difference for most game points when the empirical data is unambiguously positive or negative (i.e. the 95% interval does not include 0), as well as the magnitude of rule difference. The Uncertainty model is best able to predict the suspense difference, followed by L1 model, while almostLose model does not work as well. Note the model predictions are generated without any parameter fitting in both Figure 2.8 and Figure 2.9.

2.5.3 Discussion

In Experiment 2 we introduced a novel manipulation approach allowing us to fix the sequences presented to participants but vary their suspense implications by way of two different rules: "Bust" and "No Bust". These two rules were selected to induce different suspense trajectories even though the sequence of card draws was identical. Participants were sensitive to this manipulation, reporting different levels and patterns of of suspense depending on the condition, but the magnitude of this difference is not strong as the model predicted. This could partially be due to the suspense difference are calculated across participants (we did not allow same participant to play through identical games under two rules).

We also found that the future-belief-update model class and uncertainty-based model class more accurately captured the behavioral difference under the 2 rule conditions than the other heuristics.

Comparing to other heuristics inspired by previous studies of suspense, the "fear-of-losing" models were unable to account for the data well, but the current uncertainty-based models performed relatively well, suggesting that both current and prospective uncertainty play a role into suspense judgments. We will give an overall analysis combining all the empirical data in the final Discussion section.

2.6 Experiment 3: Manipulating willingness to play more games

After examining the factors predicting people's perceived suspense, we would like to ask a different question: what are the *effects* of suspense? This is well motivated by Ely *et al.* (2015) where suspense is hypothesized as a non-instrumental utility which people try to optimize upon.

In the context of our card game paradigm, we specifically asked: Are people who have played more suspenseful games also more willing to play more games? Previous studies have suggested that suspense levels correlates with engagement in role-playing computer games (Klimmt *et al.*, 2009), indicating a positive relation between suspense and willingness to play. Here, in a much simpler game will we see a similar effect?

2.6.1 Method

2.6.1.1 Participants

We ran 242 participants (age M = 36.3, SD = 21.5, female 123, 2 undisclosed gender). In the following analysis we will also include the 191 participants from the Experiment 1. For the 242 participants they received no bonus instruction while the 191 participants from Experiment 1 re-

ceived half bonus if they win the additional game. The participant number was expected to be around 200 due the same calculation as experiment 1 and we ended up recruiting more due to a technical error.

2.6.1.2 Procedure

Experiment 3's procedure was the same as Experiment 1 except that, after finishing all games, participants were shown the following message: "All the required games are done, thank you! However, you can also play one more game (with [half/no] bonus)". Then, participants could choose either to stop or to play one more game. If they chose to continue, a random game that they had not experienced before would be presented. We designed the two payment conditions because the monetary reward itself is a critical confounding factor for the behavior being studied, i.e. incentivising continuation orthogonally to the potential intrinsic reward of suspenseful engagement. Exploring two levels of payment helped to avoid the minor modulation of suspense condition being overshadowed by the effect of the monetary incentive.

Notice that a decision of playing more games clearly involves complex considerations: the economic return, the opportunity cost of spending the time playing, as well as any fun derived from the game. To disentangle these motivations, after the full task finishes we asked the participants what was the reason for their decision.

Our hypothesis was that participants assigned to high-suspense condition would be more likely to play one more game than those in the low-suspense condition.

2.6.2 Results

For the 191 participants in the half-bonus variant, we found a slightly higher proportion of those in the high suspense condition chose to play one more game (77 out of 96, 80.2%) than those in the low predicted-suspense variant (70 out of 95, 73.7%). Likewise, in the no-bonus condition, participants in the high suspense condition were more likely to play another game (27 out of 124 (21.7%) compared to 20 out of 118 (16.9%).

To get a Bayesian interpretation of these results, we treated the choice as a random variable drawn from the Bernoulli distribution with a bias rate r then calculated the posterior of bias rate $P(r|data)$ based on a uniform beta prior $B(1, 1)$. In the half-bonus variant, the two distributions are separated only by a 55% density interval (Figure 2.10). This is not a statistically significant result (p=0.66 in Fisher exact test). In the no-bonus variant, two distributions are separated by a density interval of 49%. The trend is also not significant in the no-bonus condition (p=0.34 in Fisher exact test).

To combine the two conditions, we performed a logistic regression on the choices with bonus

variant and suspense condition as independent variables, using the statsmodels package in python (Seabold & Perktold, 2010). The coefficient for suspense condition was 0.34 (z=1.42, p=0.16, 95% confidence interval [-0.13, 0.81]), again indicating the higher suspense increases the possibility of choosing to play one more game, though not significant statistically.

To address the concern that the suspense manipulation is not directly reflecting participants' perceived suspense, we also grouped participants by their average self-reported suspense (Figure 2.10, splitting the whole sample to two equal sized groups based on each participant's average reported suspense throughout the games. The result is again non-significant (Half bonus: 51% density interval for distribution separation; no bonus: 52%) but directionally consistent with our hypothesis.

2.6.3 Discussion

Experiment 3 provides some preliminary evidence that suspenseful experiences in a task may enhance people's willingness to engage for longer. This is is in agreement with Ely et al's conjecture that suspense function as a and some previous work in computer games (Klimmt *et al.*, 2009) where they framed a game in qualitatively different ways (with or without threat). Our manipulation of card sequence difference is much more subtle and qualitative, yet still induced certain effects.

This is a surprising effect especially given that we conducted the study online where the task platform provides very little hindrance to task switching thus relatively high opportunity cost for playing additional games with reduced monetary reward. And more importantly, workers on Amazon Mechanical Turk are most strongly motivated by payment, not for fun (Kaufmann *et al.*, 2011). Yet still, in the self reported reason of continuing play games, we saw the fun experience still accounts for participants' choice. We coded the existence of key words to represent two categories of motivation: "win/money/bonus" signaling the drive of extrinsic outcomes, "fun/enjoy" signaling the intrinsic motivation. In the no bonus variant, the majority of participants that continued reported that they did so "for fun" (21,61.8%), and fewer reported they played "to win" (13,38.2%). In the half bonus condition, although most people (81,63.3%) reported their reason as related to winning or getting more bonus or money, there was still frequent mentions of "fun/enjoy" (47,36.7%). This further hints that the fun of game will causally make people play more, and suspense could be one component of that motivation.

2.7 General Discussion

In this paper, we introduced a new paradigm for measuring suspense dynamics. Our task involved a card game designed to lead to diverse situations in terms of current and expected uncertainty. Un-

like previous studies of suspense that have relied on qualitative theories on suspense and relatively coarse manipulations[2], this new paradigm facilitates testing theories of suspense in a quantitative way. Specifically, we tested the Ely et al. (2015) proposal under which suspense is driven by expected future belief change. This theory successfully captured the behavioral data across a range of card sequences (Experiment 1) and captured differences between two game rules for the same card sequence (Experiment 2). Overall, the class of forward looking models shows better generalizability than the alternatives we considered.

2.7.1 Evaluation of models across all experiments

We now combine all the empirical data we have from Experiments 1, 2 and 3 to evaluate the models. Results are summarized in Table 2.4. We fitted the same group noise parameter across all data for each model. Detailed description of the data and comparison are described in the appendix.

The overall winner is the "future-belief-update" using the L1 norm. In terms of correlation with the response averaged across participants, all the "future-belief-update" models have higher correlation coefficients than the other heuristics. In terms of likelihood fitting of each individual's response, the two "future-belief-update" models in L1 and Hellinger forms fit the best, followed by Uncertainty model. The heuristic of "fear-of-losing" performs worst overall. Note that the "unceratinty" heuristics may perform similarly to the "future-belief-update" models in some experimental conditions but not others. Conceptually, this is understandable because the "future-belief-update" models already incorporates part of the "uncertainty" model: these models make more and more similar prediction as the time horizon approaches the end of the task. It is only at the beginning of the game that "future-belief-update" models will predict lower suspense than "uncertainty" model." It is possible that better resolution of these theories might come from larger experiments with a greater variety of stimuli.

Table 2.4: Model comparison for all Experiments.

	Future belief update				Uncertainty		Fear of losing	
	L1	Hellinger	L2	ΔH	uncertainty	cardDiff	almostLose	pLose
Pearson's r	0.82	0.81	0.74	0.74	0.69	0.60	0.65	0.50
Likelihood difference	8.78	7.89	5.92	5.77	6.20	3.85	3.22	0.9

Note: Columns and rows as in Tables 2.2 and 2.3.

Across our three experiments, we thus have good evidence that an anticipation based "future-belief-update" model captures something key to human suspense patterns that cannot be captured

[2]While Comisky & Bryant (1982) do take a somewhat quantitative approach, their manipulation was not entirely numerical. Instead, manipulated text was used qualitative language — "the chance was...": "...absolutely nil" / "...extremely slim at best" / "...somewhat against" / "...roughly even" / "...totally certain"

by present focused uncertainty based models or simpler heuristics. Yet still, we had the concern that people may use a combination of heuristics instead.

Thus, we finally tested a hybrid model allowing the linear combination of the heuristic models of "uncertainty", "cardDiff" and "almostLose". With linear regression against the aggregate response, the hybrid heuristic model gives slightly higher Pearson's r (0.83) than the L1 model (0.82). For the likelihood fitting, however, the hybrid heuristic model gives likelihood 7.68($*10^{-}2$) per trial, which is smaller than the single model of L1 ($8.78*10^{-}2$ per trial) and Hellinger ($7.89*10^{-}2$). Given that the "future-belief-update" model generates suspense prediction without any parameter tuning, this seems to be an empirically promising account of suspense.

2.7.2 Limitations and future directions

This line of experiments has some limitations that will necessitate future investigation. First, although the best model we found was in the family of future belief change models, there is no clear explanation why L1 belief distance metric is consistently preferred. By comparing the different metrics in Appendix Fig 2.14 and 2.13, we found that all the other metrics underestimates the suspense compared to the L1 norm. Whether this preference for belief update judgement relates to other belief-related psychological phenomenon is worth exploring (e.g., does the L1 measure predict judgments of surprise?).

Second, it is likely that participants' understanding of suspense was somewhat heterogeneous and that their subjective access to this quantity was limited. That is, self reports may have been driven by arousal responses from a mixture of sources, including anxiety elicited by uncertainty and fear elicited by potential failure as well as expectations about future belief change. The belief change model class could capture some of these sources while heuristic models may have captured other valenced sources such as fear of losing. Further studies that manipulate outcome valance (either with wins being rare or common, and likely or unlikely given a game state) may help to disentangle these differences.

Third, given the subjectivity of suspense, we used self report. However, self report constantly interrupts the flow of experience. We hope this work could serve as a foundation for further experimental work utilising a more implicit measurement of suspense, for example, EEG signal of suspense.

Fourth, while we took care to construct situations with a wide range of suspense levels, we were still restricted to the space of single-player card games where many potentially important factors of suspense are absent, presumably limiting the ceiling levels of suspense participants experienced. Future work could try to study settings that are known to be particularly captivating yet still simple enough to enable computational analysis. Alternatively, more sophisticated computational tools

can be employed to process more complex information (for example, Wilmot & Keller 2020 used natural language processing techniques to estimate suspense in short stories).

Last, although in Experiment 3 we found evidence suggestive that suspense might increase people's willingness to further engage in a task, the effect was very small and not statistically significant even in our substantial sample size. Besides increasing the range of suspense, we may be able to find more sensitive measurements of the effect of suspense, such as modulation of attention (see Bezdek *et al.* 2015) or willingness to pay to see an outcome. These steps will help provide a normative account of why expected future belief change produces the physiological state that we call suspense.

2.8 Appendix

2.8.1 Stimuli design

In experiment 1, we first simulated 5000 games, then filtered out games that looked "fake", as was complained in pilot studies. Specifically, we excluded games that 1) have certain cards appear $>$ 2 times or 2) have a card pair being selected > 1 times. For the rest of "valid" games, we aimed to present stimuli with different level of suspense. Pilot study suggests that both the Ely et al. suspense and the "pLose" model predict empirical suspense well, we picked a "hybrid suspense" that averaged over these two model predictions as our index. Specifically, we selected the lowest 10% and highest 2% suspense games into our final round of game candidates, including winning and losing games. This asymmetry comes from the fact that for losing games, it was hard to elicit low suspense, thus more games are included for the "low predicted-suspense" group. In our round of simulation, there are 59 low-win, 13 low-lose, 46 high-win, 43 high-lose games. Finally, we randomly selected the number of stimuli we needed for the experiment.

2.8.2 Data from three experiments

Here we delineate the differences of three experiments' suspense report.

The game rules are same in Experiment 1 and 3, but different from Experiment 2. In Experiment 1 and 3, participants each played 3 gambling games, reporting suspense at at most three points, thus contributing 9 data point per person. In Experiment 2, each participant played 5 gambling games, each reporting 3 points.

For likelihood fitting of individual responses, we have 2139 data points from experiment 1, 1879 from experiment 2 and 2367 from experiment 3. For the regression of suspense report across participants, we have 88 game points from experiment 1 (on average 40 participants per point), 30 from experiment 2 (on average 78 participants per point) and 154 from experiment 3 (on average 10 participants per point). Therefore, the cross-participant reports may have unfairly weighted the data from experiment 3 which has more points but also significantly more noise.

2.8.3 Results from Experiment 3

The comparison of models and data are listed in table 2.5.

2.8.4 Computing model likelihood

Given a participant suspense rating R_t on a given trial t (where $R_t \in \{1, 2, 3, 4, 5\}$), we would like to obtain the likelihood of R_t given the predicted suspense of the model $\mathcal{S}_t$ (i.e. $p(R_t|\mathcal{S}_t)$).

Table 2.5: Experiment 3: Model Comparison

	Future belief update				Uncertainty		Fear of losing	
	L1	Hellinger	L2	ΔH	uncertainty	cardDiff	almostLose	pLose
Pearson's r	0.81	0.80	0.73	0.73	0.65	0.52	0.68	0.59
Likelihood difference	7.60	6.43	5.06	4.81	3.85	1.69	3.08	2.84

Note: 1st row: Correlation coefficient between the parameter-free model prediction and the raw participant report of suspense. 2nd row: Likelihood fitting all individual trials minus the baseline (random report), averaged to per trial, in the unit of 10^{-2}. All the heuristics are worse than the "future belief update" models. The best model is the "future belief update' models of L1 and Hellinger form.

We treated the response as a multinomial distribution parameterized by $\mathcal{S}_t$. To determine the exact value of this multinomial distribution, we constructed a beta distribution with its mode being the same value as the model suspense $\mathcal{S}_t$ for each game point. Then, we calculated the cumulative probability density under each 1/5 percentile of this beta distribution, mapping into the integer output of suspense, thus obtaining the likelihood function. An intuitive example of this process is shown in Figure 2.11.

Mathematically, the multinomial can be defined as follows:

$$p_k = \int_{(k-1)/5}^{k/5} \mathrm{Beta}(x;\alpha,\beta)\mathrm{d}x, \mathrm{for} k = 1, 2, \ldots, 5 \tag{2.8}$$

whose beta parameters are defined such that the mode of beta distribution is equal to the suspense prediction (scaled to [0,1]):

$$\mathrm{Beta}(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \tag{2.9}$$

where

$$\alpha = A * S + 1, \beta = A * (1 - S) + 1 \tag{2.10}$$

$A \in [0, \infty)$, is a positive free parameter controlling the flatness of the beta distribution (thus the randomness of the multinomial distribution).

For aggregate data we fit this model by minimizing the negative log likelihood with A determined by minimize function of scipy package from random starting points. We defined a baseline model where all p_k are equal across the rating options (i.e., modeling each rating being randomly selected). We compared the maximum likelihood of the suspense model to the baseline model and average over all individual card draws to present the log likelihood improvement per draw.

2.8.5 Derivation of belief update

In Equation 2.1, we used the relationship $\mathrm{E}_s[\mu^s_{t+1}] = \mu_t$, where s is a game state (i.e. the card score in our context), t is the current time point, μ^s_t is the belief of winning probability given at time t. In other words, the prior belief of winning equals the expectation of the posterior. This is part of the definition of a belief martingale (see Williams (1991), chapter 10) a universal property of Bayesian learning. In Ely et al. 2015 they directly used this property in their definition of suspense.

For an intuitive understanding, consider the following. Imagine an observer has a belief μ concerning the probability of winning a game. We can start from the end of the game then work backwards. Assume under a rule set, there are final states s_w that results in winning and s_l for losing. Thus at the final time step T, the belief in either winning or losing is just determined by the outcome of the game:

$$\mu^s_T = \begin{cases} 0, s \in s_l \\ 1, s \in s_w \end{cases} \tag{2.11}$$

In other words that the final step of the game the belief is just if the person wins or loses. Next, for one time step back at $T-1$ at state s', the winning probability would be marginalizing all the possible next steps weighted by the transition probability:

$$\mu^{s'}_{T-1} = \sum_{s \in s_l} p(s|s') \cdot 0 + \sum_{s \in s_w} p(s|s') \cdot 1 = \sum_s p(s|s')\mu^s_T = \mathrm{E}_s[\mu^s_T] \tag{2.12}$$

Figure 2.12 shows the state transition of the last step, where four possible states are included and marginalized to calculate $\mu^{s'}_{T-1}$. Similarly, the belief μ at any time t, state s is derived from the the belief at time $t+1$:

$$\mu^{s'}_t = \sum_s p(s|s')\mu^s_{t+1} = \mathrm{E}_s[\mu^s_{t+1}] \tag{2.13}$$

Note here we write the term μ_t in Equation 2.1 explicitly as $\mu^{s'}_t$.

2.8.6 Alternative models

2.8.6.1 Formulation of the Heuristic models

First, for the heuristic of "uncertainty", if people keep track of a probability of winning, the uncertainty should be the highest when the probability of winning is 0.5 and lowest when it is 0 or 1. To capture this idea, we use the entropy of the belief distribution:

$$S_{\text{uncertainty}} = \mathrm{H}(p_t) \tag{2.14}$$

Alternatively, instead of keeping track of winning probability which requires a full simulation towards the end of the game, people may instead only be concern about how much uncertainty they have now given the 2 candidate cards, or simply, the difference between the two cards:

$$S_{\text{Carddiff}} = |v(1) - v(2)|/(v_{max} - v_{min}) \tag{2.15}$$

where v denotes the value of single card, normalized by the maximum card value difference given all possible values of the cards.

We also tried an alternative normalization using the maxixmum possible difference given the current deck (thus different in every game), but this model does not perform as well.

Secondly, for the heuristic of "fear of losing", if people keep track of a probability of winning, then this can be defined as:

$$S_{\text{pLose}} = 1 - p_t \tag{2.16}$$

In this formulation of pLose model, maximum suspense is achieved when the game is definitely losing. Alternatively, when people are definitely losing, they may instead no longer feel suspense, expressed as almostLose model:

$$S_{\text{almostLose}} = \begin{cases} 1 - p_t, \text{if} p_t > 0 \\ 0, \text{if} p_t = 0 \end{cases} \tag{2.17}$$

Another possible formulation is that people can approximate the closeness to losing by how far is the largest of the two cards drawn from the deck is from the bound:

$$S_{\text{toBound}} = \text{Dist}(\text{Sum}_t + \max(v(1), v(2)) - \text{Bound}) \tag{2.18}$$

Where Sum denotes absolute value, v denotes the value of the next coming card. However, it is very unclear how to normalize this distance. We tried using linear normalization and negative exponential with different constants but none will fit all the games better than the random guess, even after adjusting the constant according to the rules. We deemed this model as an ill-defined model and thus did not report it here.

2.8.6.2 Intuition of the difference between different belief update norms

We here show the magnitude of different belief update metrics given the current probability of winning ($p(now)$) and a potential next step belief ($p(next)$), see Figure 2.13. Note that some combinations are impossible since the expected future belief should be equal to $p(now)$.

More intuitively, we show that when the current belief is 0.5 (i.e. maximally uncertain what

will happen next), the different metrics will give suspense as in Fig2.14.

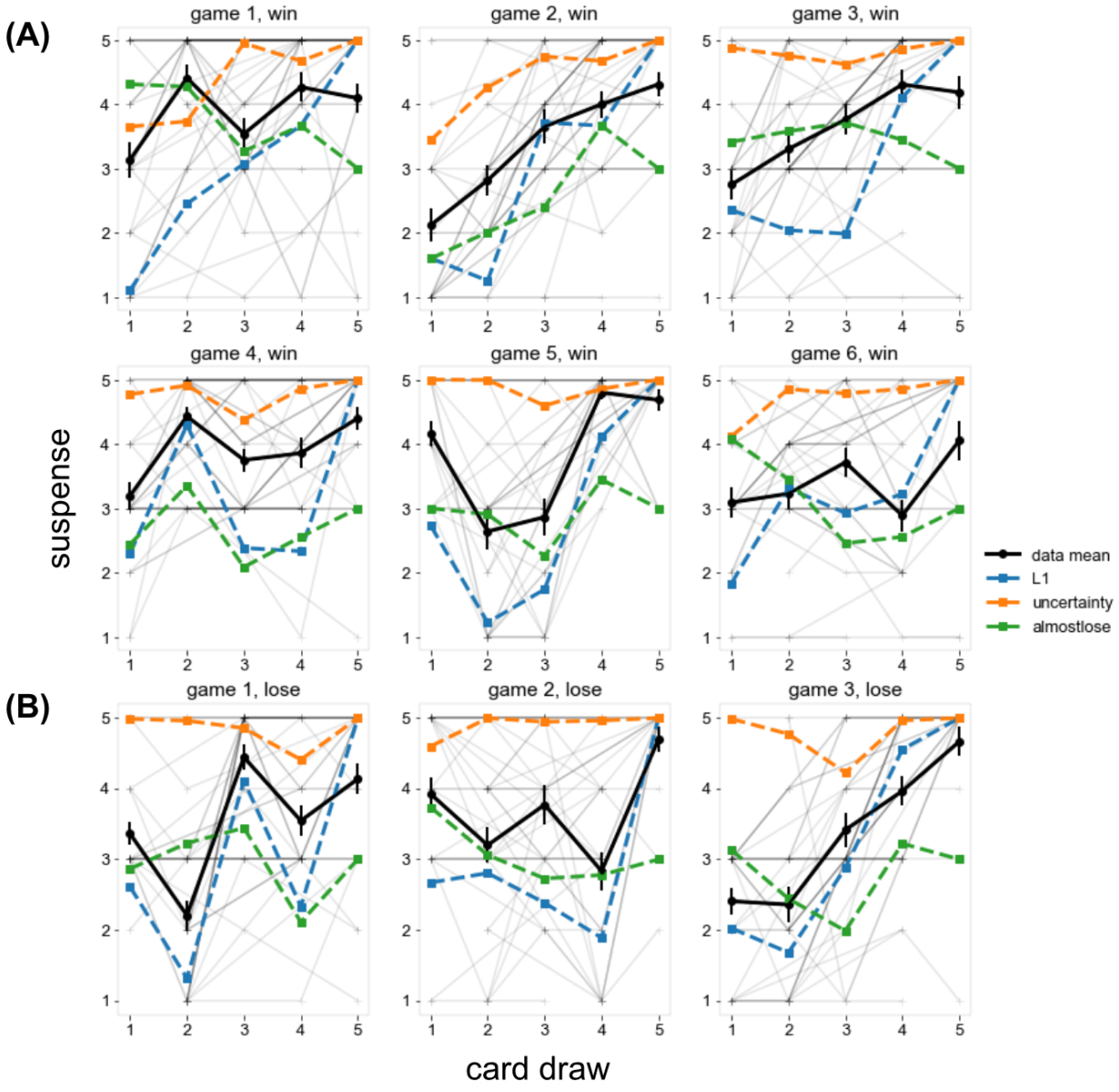


Figure 2.4: Experiment 1: Average suspense judgments and model predictions across all high predicted-suspense games. (A) games resulting in a win; (B) games resulting in a loss

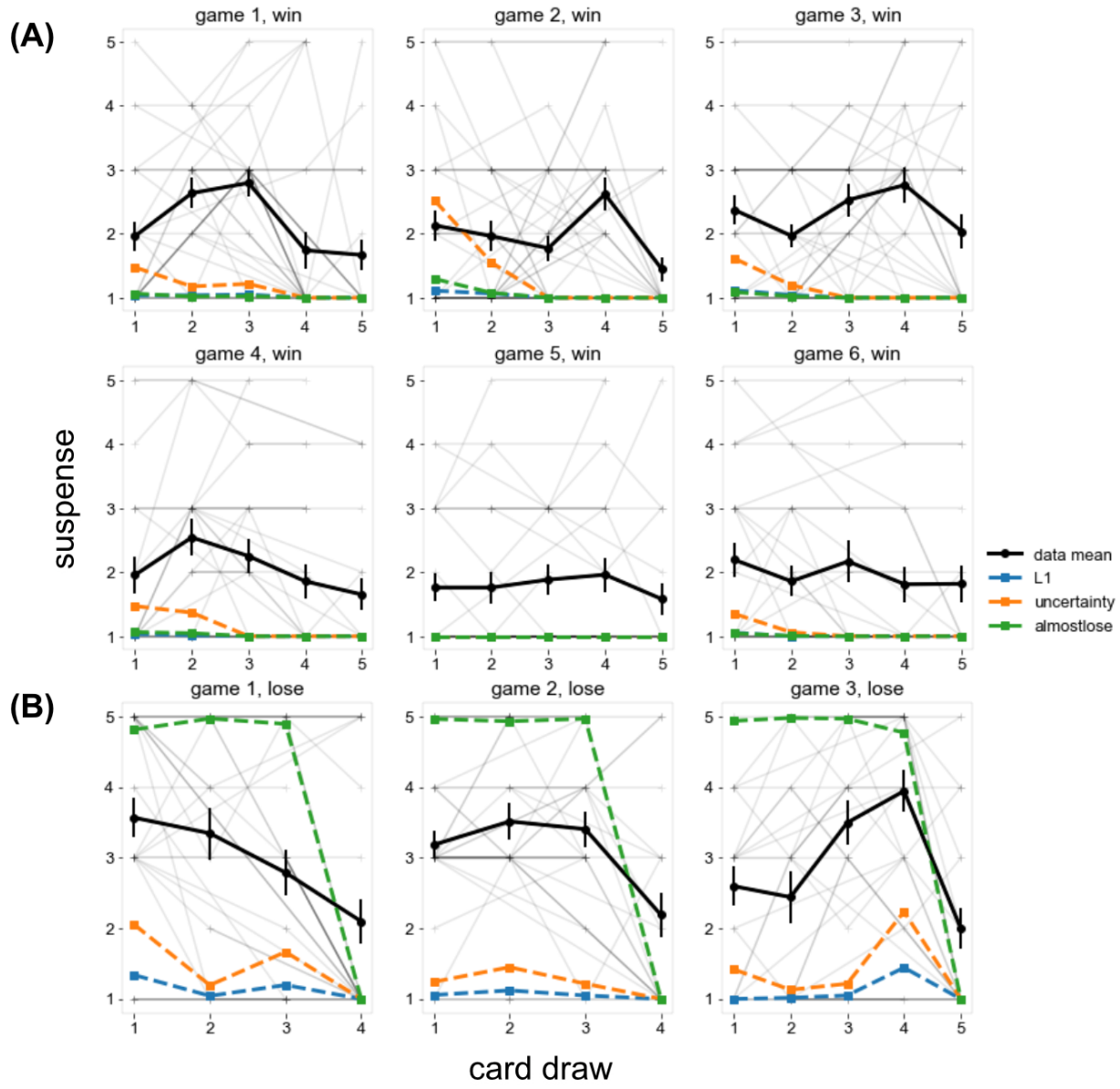


Figure 2.5: Experiment 1: Average suspense judgments and model predictions across all low predicted-suspense games. All the models show systemic biases here. (A) games ended up winning; (B) games ended up losing

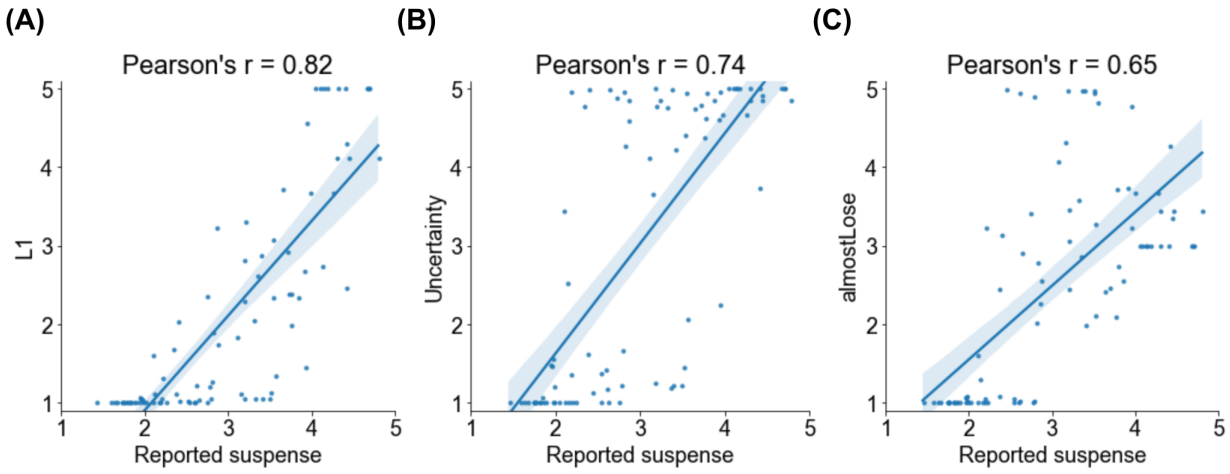


Figure 2.6: Experiment 1: Correlation between model predicted and averaged participant reported suspense. A. "Future belief update" model with L1 norm. B. Uncertainty model C. " Fear of losing" model

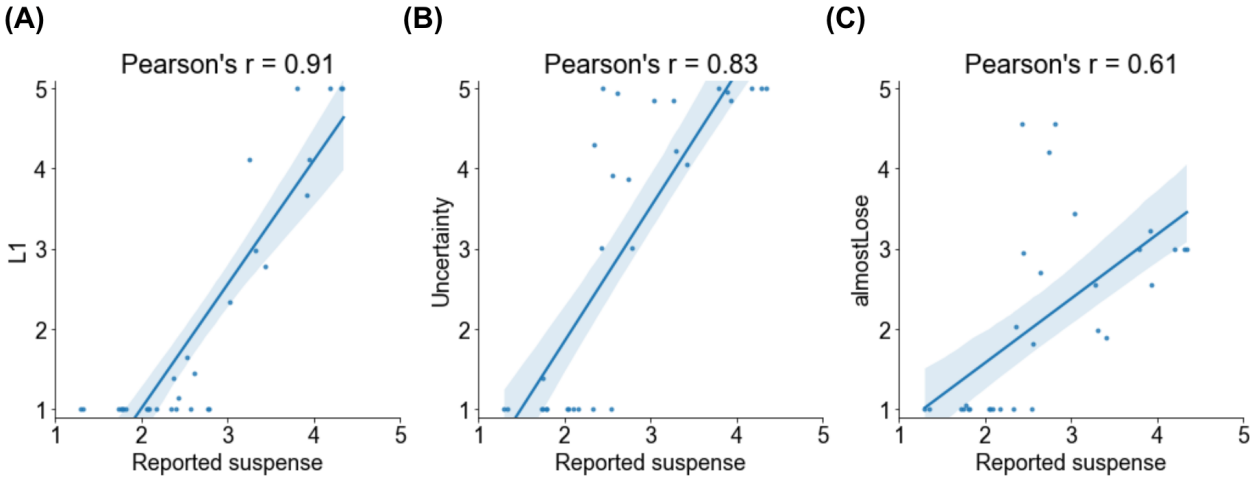


Figure 2.7: Experiment 2: Model predicted and average judged suspense for the best-fitting model from each conceptual class. Games of two rules are all included. The best model from each conceptual category are shown: A. "Future belief update"; L1 norm. B. Uncertainty heuristic; current belief uncertainty. C. "Fear of losing" heuristic; (almostLose)

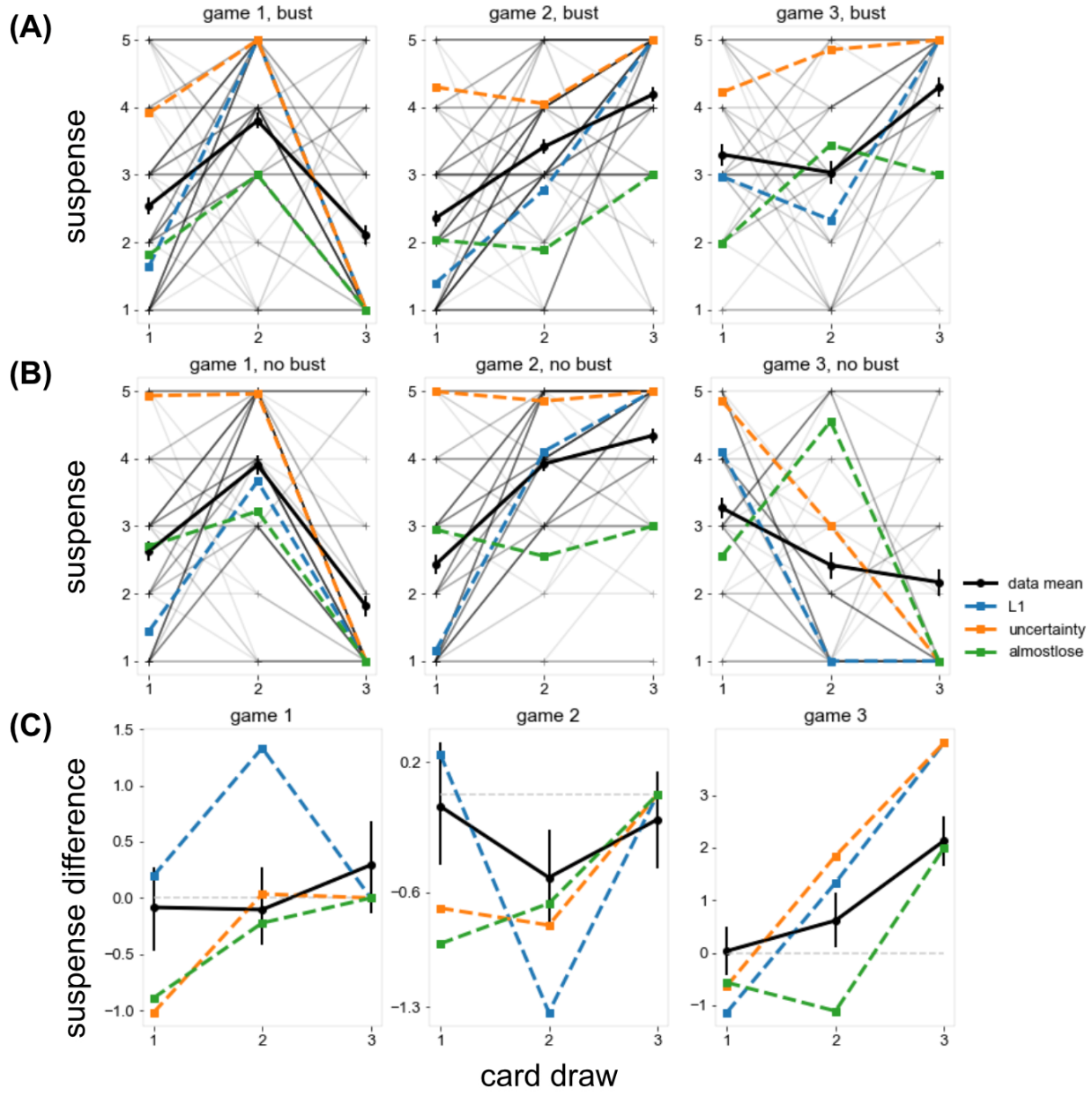


Figure 2.8: Experiment 2: Raw and relative suspense judgments and model predictions across rule conditions for all the games. (A) "Bust" condition. (B) "No-bust" condition. (C) Inter-condition difference ("no-bust" − "bust").

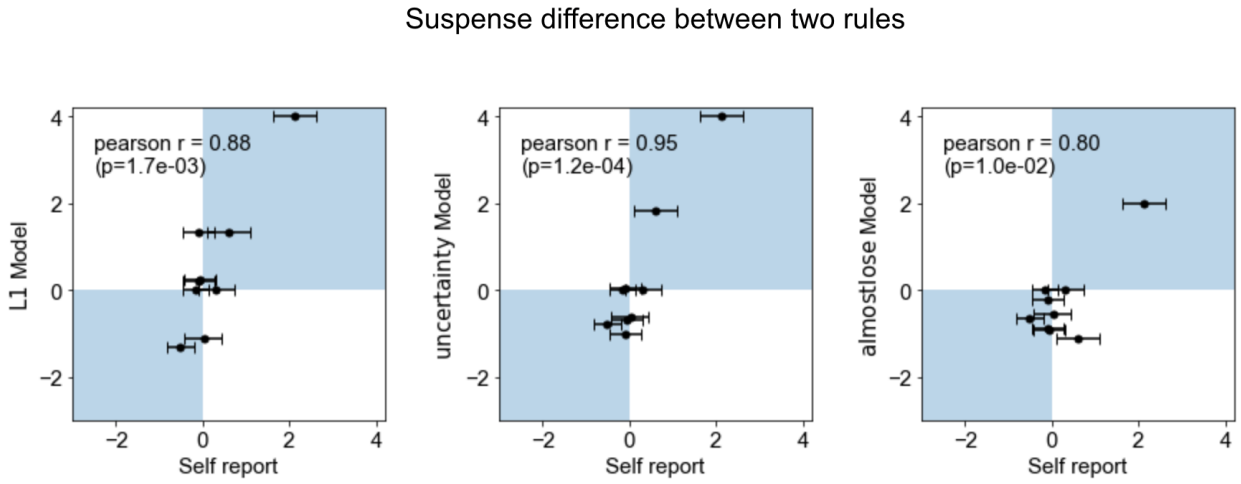


Figure 2.9: Experiment 2, Model predicted and actual suspense difference between "Bust" and "No bust" rule conditions. Each point represents one game point under the two rules. If the model predicts the rule-induced difference to be the same direction of empirical data, the data point will appear in the 1st and 3rd quadrants (shaded in blue). Errorbars on x-axis show bootstrapped 95% confidence intervals (calculated by randomly choosing the ratings between two conditions of the same game point, taking the rating difference and repeating 1000 times. Then the middle 95% of this 1000 samples becomes the confidence interval).

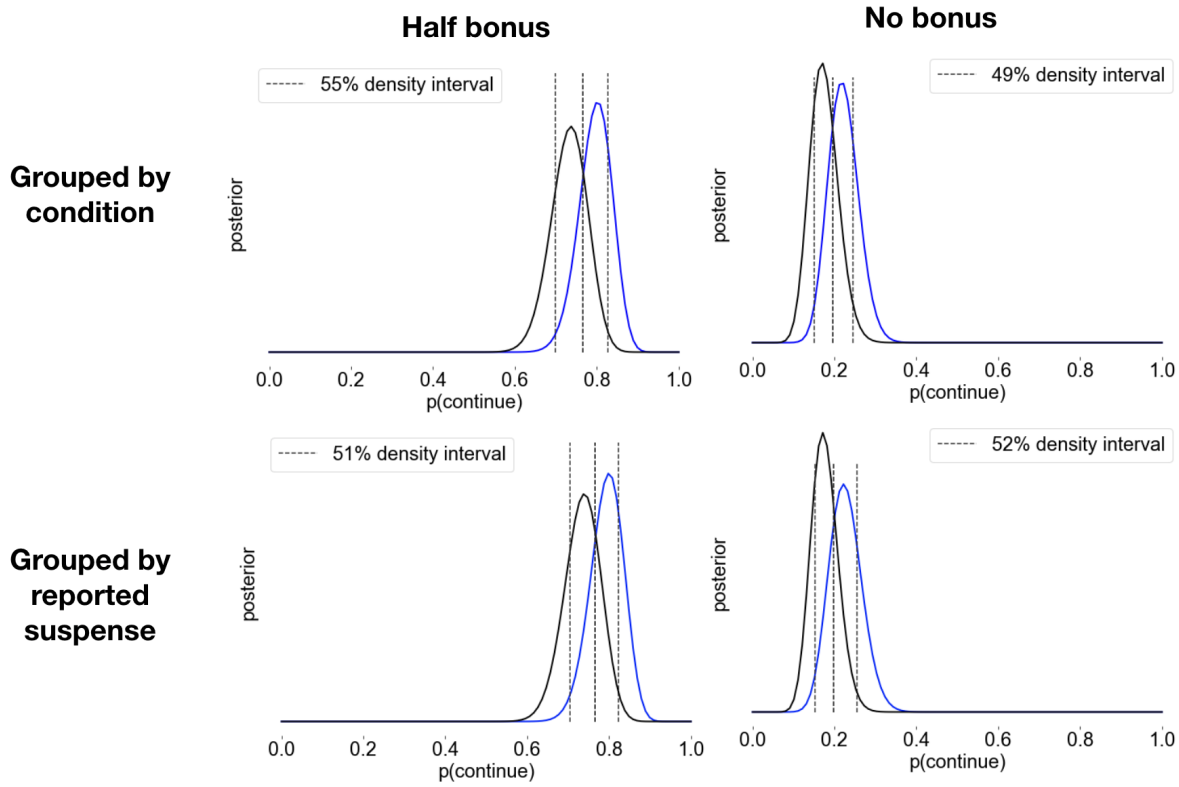


Figure 2.10: Experiment 3, Bayesian posteriors for condition-dependent bias terms. The density interval marks the boundary separating the two distributions. The first row we separated the two distribution by the assigned high/low suspense condition; The second raw we separated participants by whether their self-reported suspense falls into the 50% percentile among the whole sample.

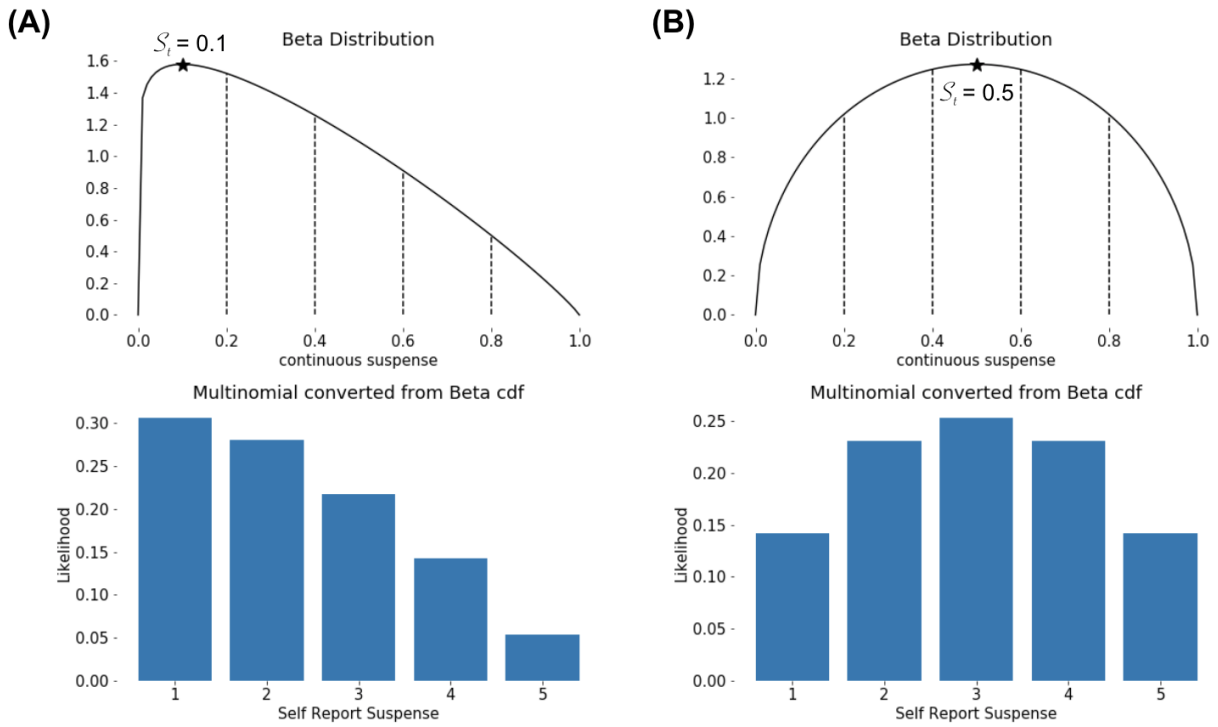


Figure 2.11: Converting the continuous model suspense output to the discrete self report. (A) The model suspense is 0.1, which in turn determines the beta distribution with a mode of 0.1, as marked as the star. Then the area under each 1/5 percentile of this beta distribution (denoted by the break lines) is converted to the value of multinomial for choosing each suspense output. Since model prediction is quite low, in the self reported suspense of 1 being most likely response. (B) With model suspense of 0.5, the the area under each 1/5 percentile of this beta distribution (denoted by the break lines) is symmetric with the highest being the middle percentile. Since model predicts a medium level of suspense, the corresponding multinomial function gives a highest likelihood of responding 3.

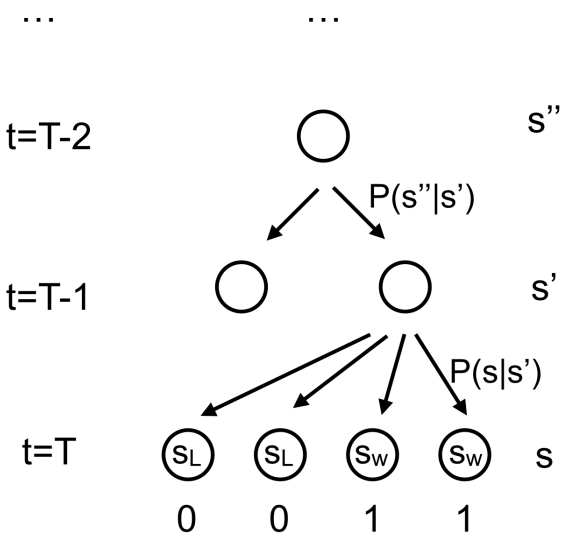


Figure 2.12: Illustration of the last three time steps of state transitions.

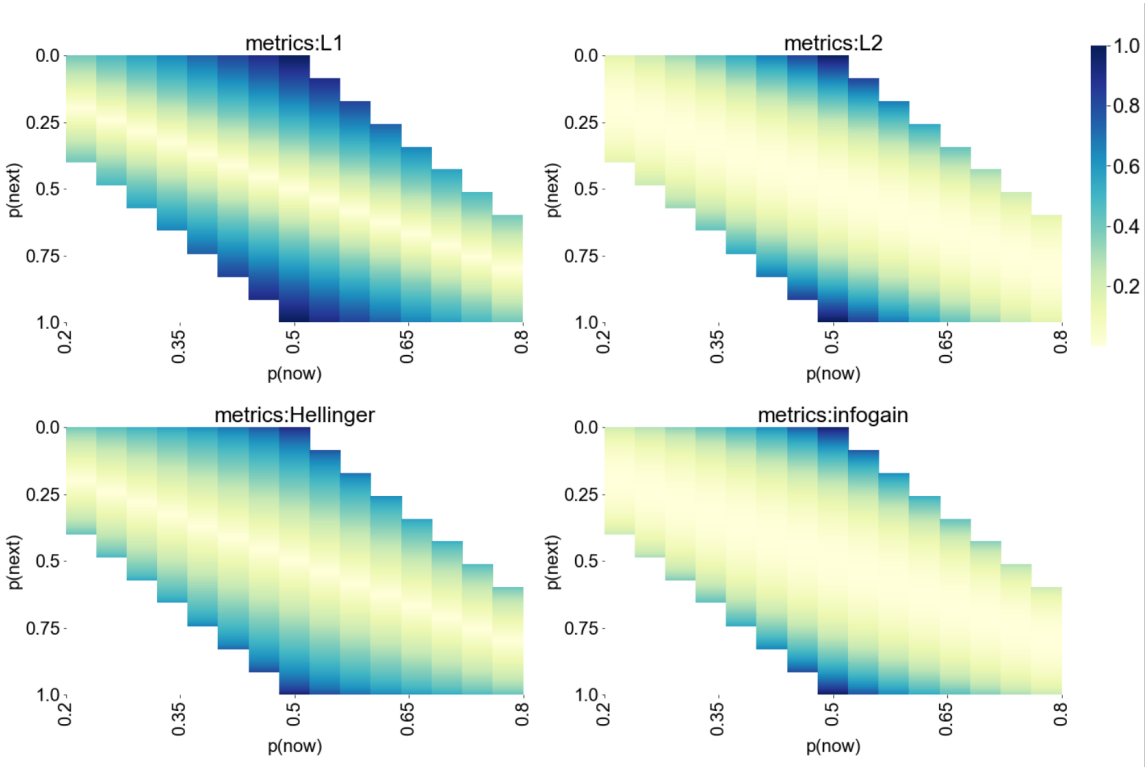


Figure 2.13: Comparing different metrics.

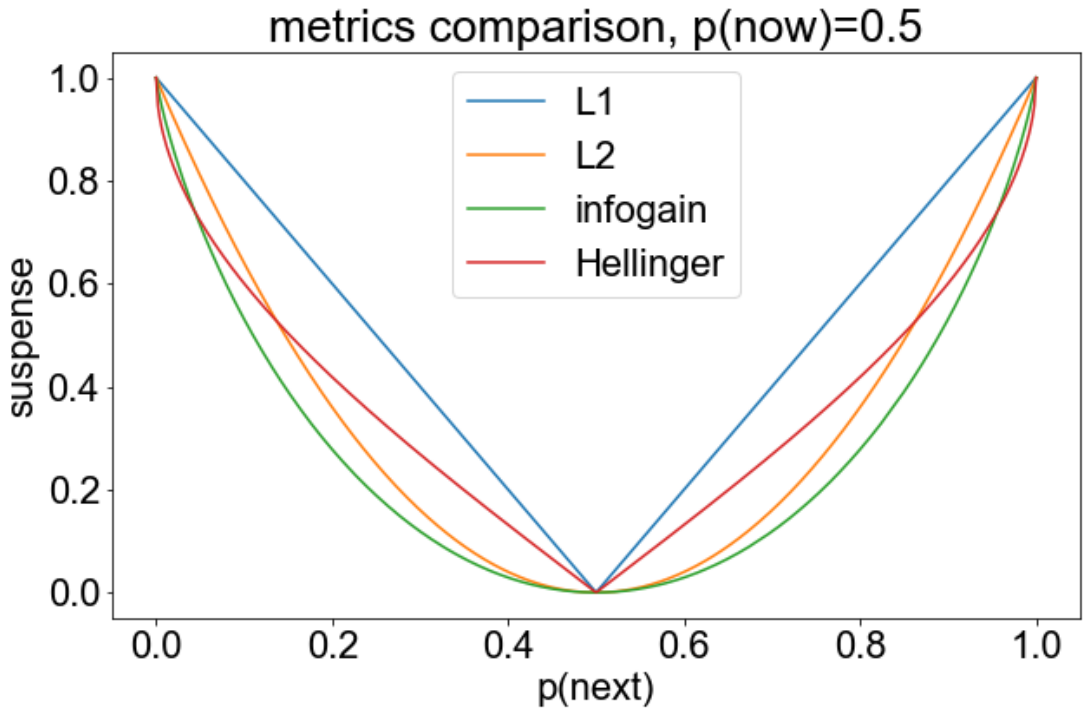


Figure 2.14: Comparing different metrics given the special case of p(now)=0.5

CHAPTER 3

Between simple and complex: Good explanations include (only) the strongest causes

3.1 Abstract

People have an intuitive feeling about 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 than 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), 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 could there be more fundamental principles underneath these individual explanatory virtues??

Among all the explanation virtues, the factors of simplicity and complexity have 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 beyond scientific practice, it can 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 the conditions making people prefer either a simple or complex explanation (defined in terms of the causal structure, most often the number of causes, but with 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 preferences shift towards simplicity or complexity. However, these kinds of methods usually only address the simplicity preference or the complexity preference separately. Looking into related manipulations is helpful in exploring different motivations for either simplicity or complexity preference, but this method could not answer the following question: is there a single mechanism or algorithm that could unify these phenomena, 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 against 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 equates to a better chance of generating such an 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 expressed 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 the likelihood is always 1, i.e., the causes are deterministic), then they 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 add up to be higher (we will explain this mathematical intuition in the Method section in Experiment 1). However, none of these studies

are actually formulated in terms of posterior probability, nor 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, we have two goals for the current study: 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 of choosing between simple versus complex explanation, we will 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.

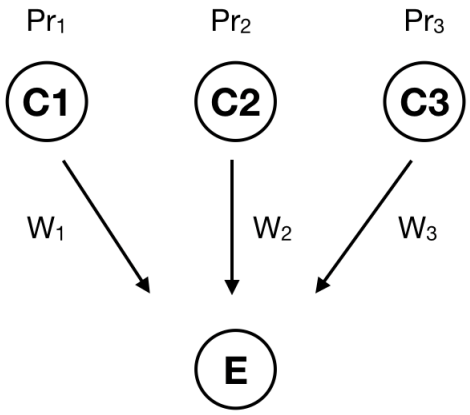


Figure 3.1: We use the common effect structure (also called collider structure) for our experiments. Causes C1, C2 C3 are independent from each other, with a prior probability Pr_1, Pr_2. Pr_3. Each cause probabilistically cause the effect E, with the probability $P(E|C_i, \neg C_{j\neq i})$ denoted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

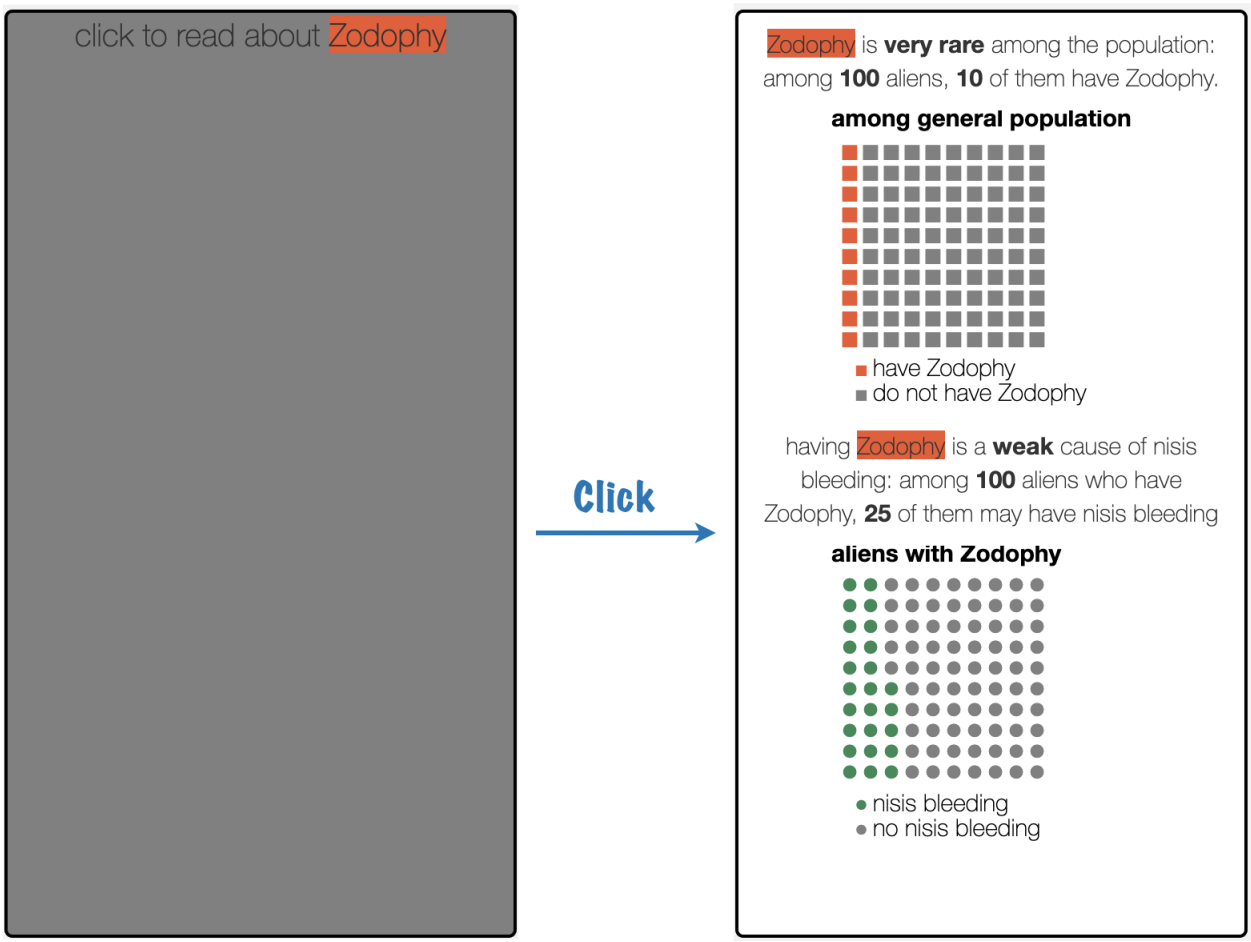


Figure 3.2: Example of the information card provided for a given cause. Initially the blank card (left) is shown on the screen. After participants click on the card, it flips and show the information card (right). On the information card, both the prior and causal strength information are presented in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

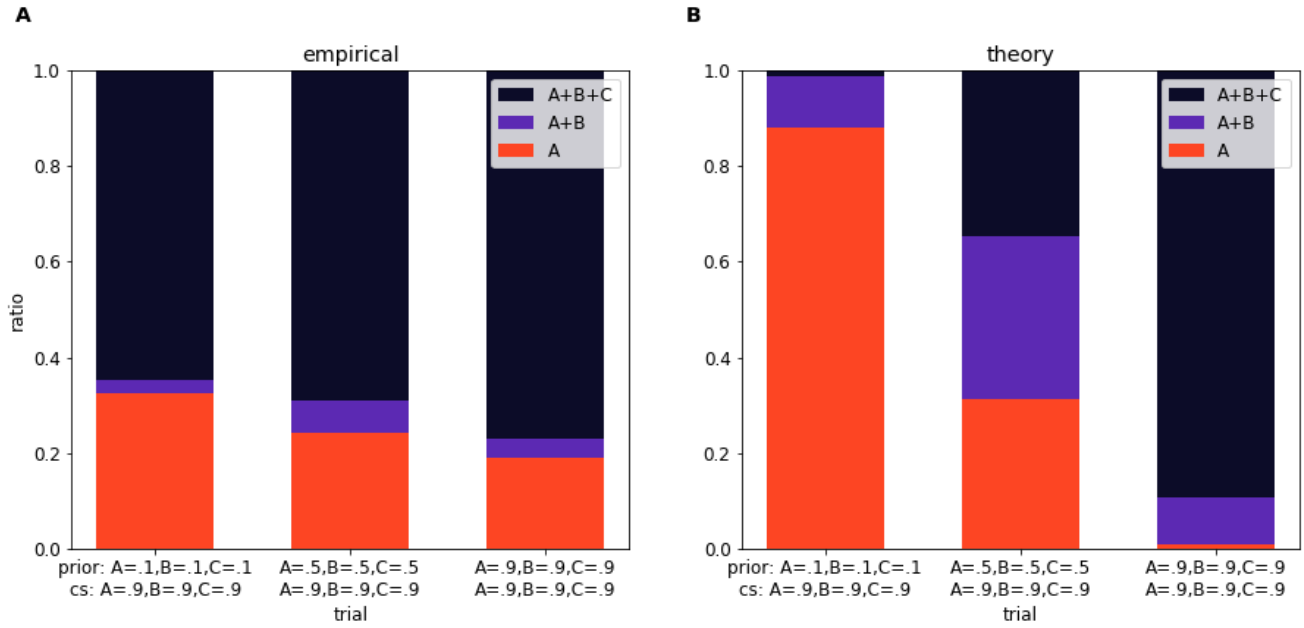


Figure 3.3: Choice ratio for trials with different prior and same causal strength. Left panel is the empirical average across all participants. Right panel is the theory predicted ratio based on the Bayesian posterior model.

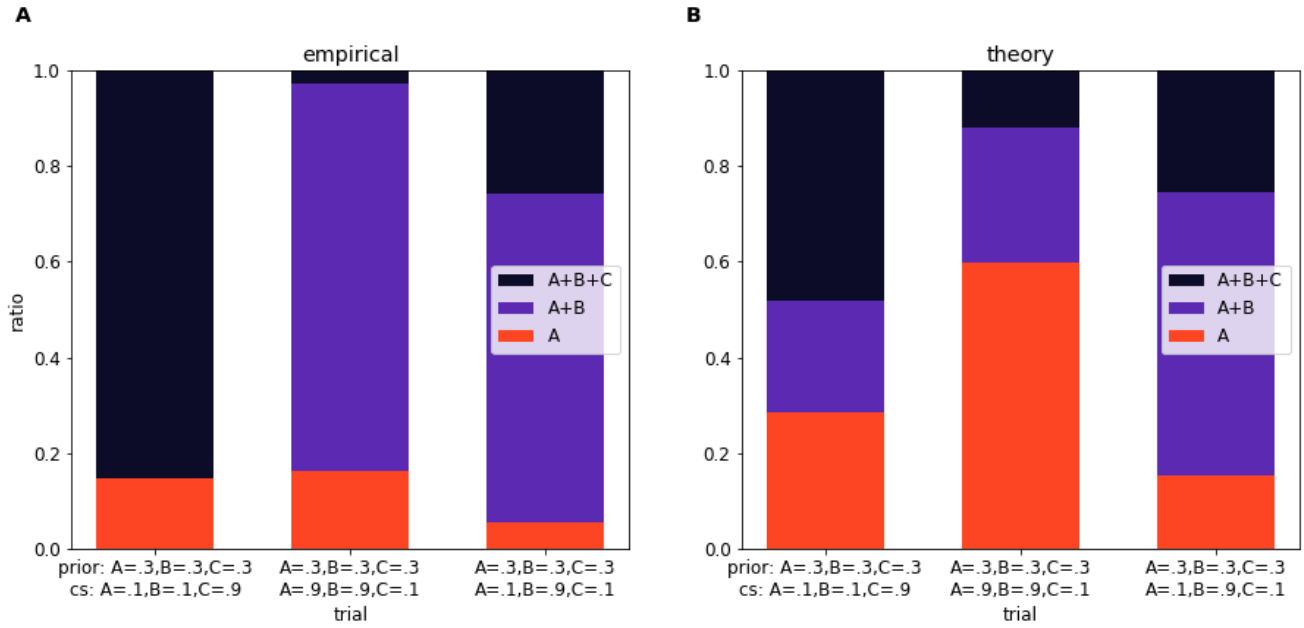


Figure 3.4: Choice ratio for trials with different causal strength and same prior. Left panel is the empirical average across all participants. Right panel is the theory predicted ratio based on the Bayesian posterior model.

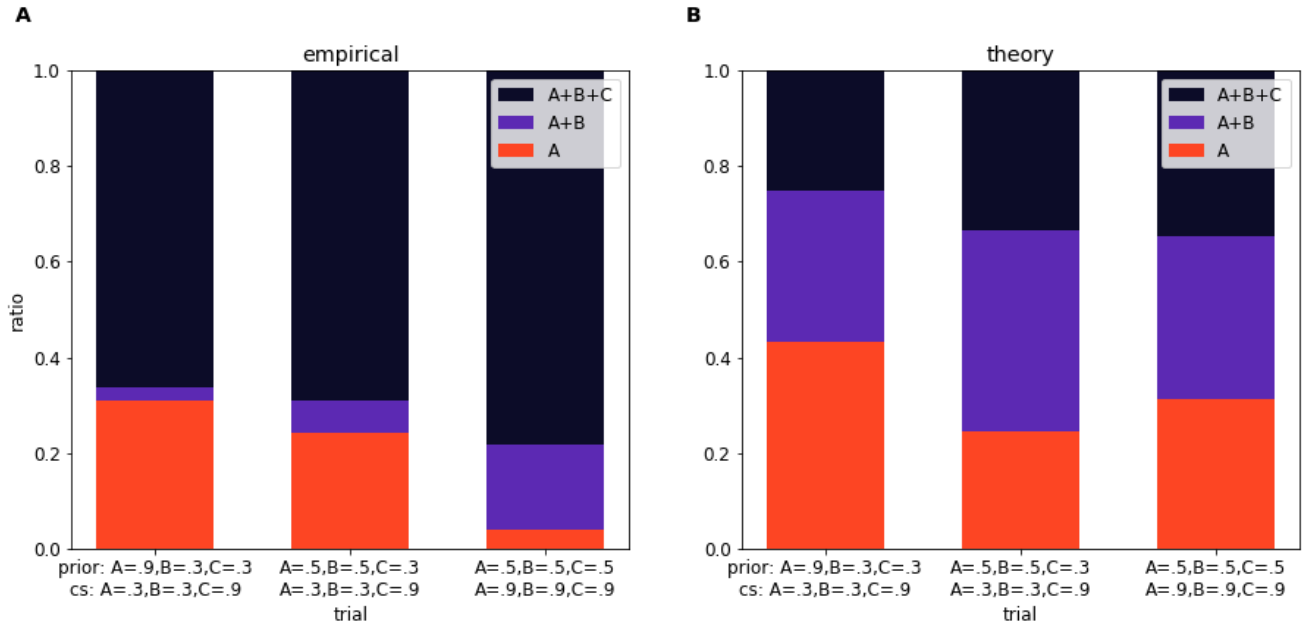


Figure 3.5: Choice ratio for trials with similar level of posterior according to the Bayesian model. Left panel is the empirical average across all participants. Right panel is the theory predicted ratio based on the Bayesian posterior model.

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	0.9, 0.9, 0.9	A+B+C	A+B+C (64.9%)
1	0.9, 0.9, 0.9	A+B+C	A+B+C (68.9%)
2	0.9, 0.9, 0.9	A+B+C	A+B+C (77.0%)
3	0.1, 0.1, 0.9	A+B+C	A+B+C (85.1%)
4	0.9, 0.9, 0.1	A+B	A+B (81.1%)
5	0.1, 0.9, 0.1	A+B	A+B (68.9%)
6	0.3, 0.3, 0.9	A+B+C	A+B+C (66.2%)
7	0.3, 0.3, 0.9	A+B+C	A+B+C (78.4%)

Table 3.1: Comparing the predicted best explanation with empirical average best explanation. The causal strength columns list the values in the order of A, B and C. 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	0.1, 0.1, 0.1	0.9, 0.9, 0.9	A, B and C (64.9%)	A, B and C	only A
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
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	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. The prior and causal strength columns list the values in the order of A, B and C.

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.

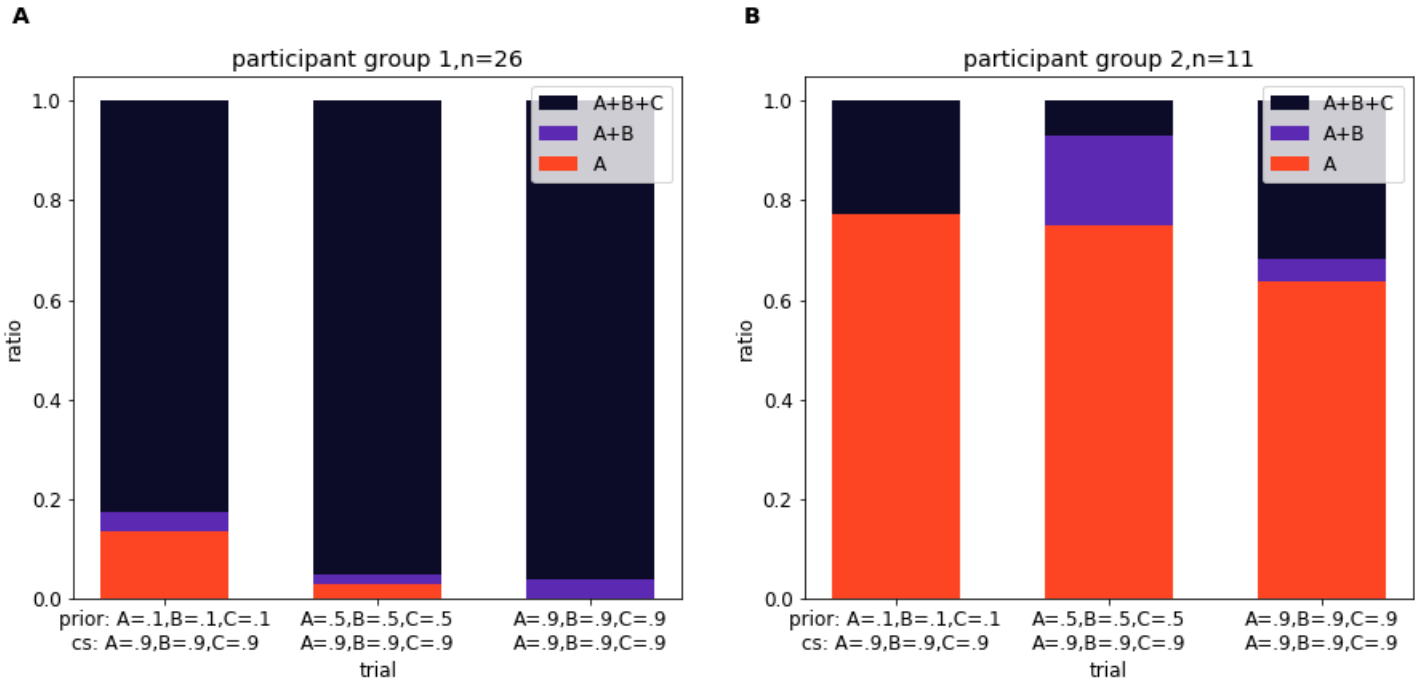


Figure 3.6: Choice ratio of the three explanations for trials with equal prior and causal strengths, separated by the two clusters of participants.

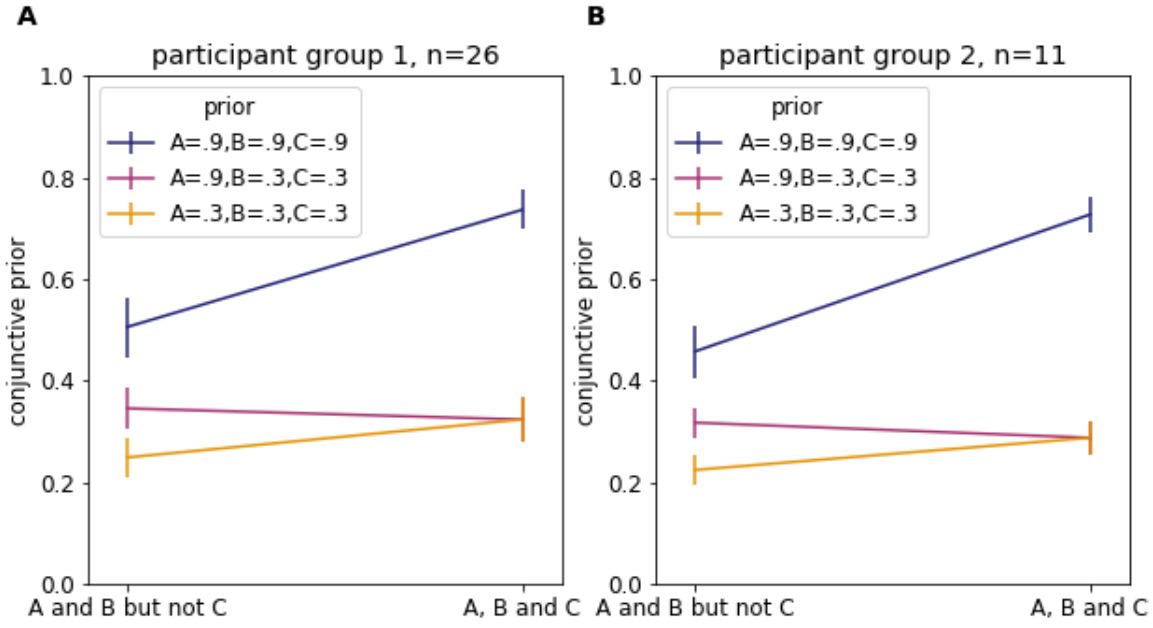


Figure 3.7: Rating of conjunctive priors, separated by the two clusters of participants.

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 fails 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

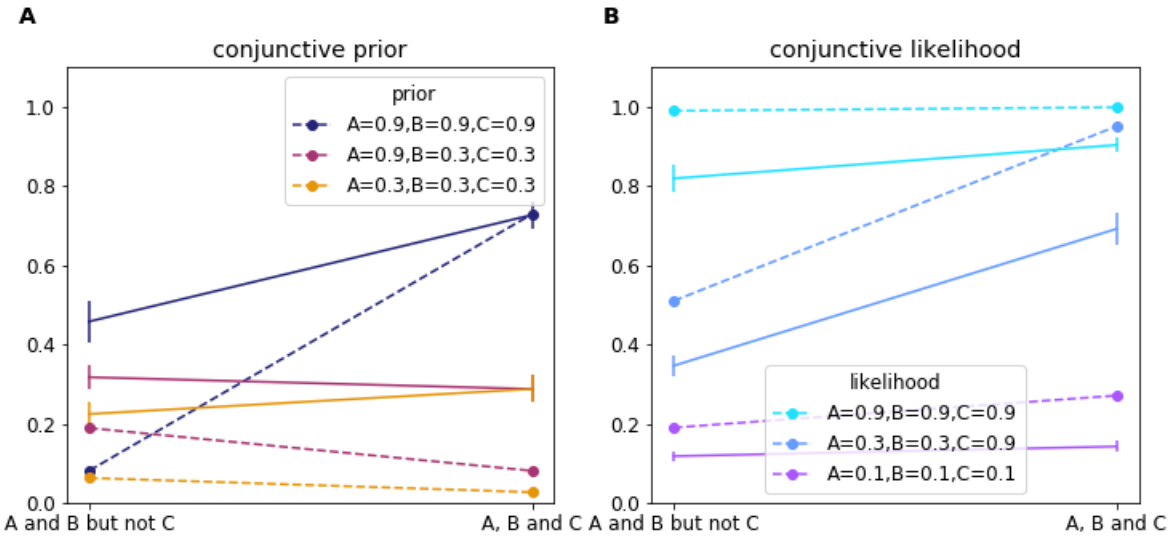


Figure 3.8: Contrasting the empirical data (solid line) with theory prediction (dash line) regarding the prior and likelihood (causal strength) of conjunctive causes. The error bars on solid lines represent standard error. Note that participant reports were in the range of 0-100 and here we normalize 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. People may treat deterministic causes differently than probabilistic causes (Johnson *et al.*, 2019), 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

We started analyzing the behavioral trends by separating the trial types because the trial type itself is a biggest factor for different explanation preferences.

For the trials with identical causes, despite both the prior and causal strengths have different values than Experiment 1, the explanation preference is very similar (Figure 3.9; compare with

Figure 3.3 in Experiment 1). Specifically, in all of those trials, the explanation "A+B+C" is the most preferred explanation overall, indicating complexity preference. But there is also evidence for simplicity preference in that the 2-cause explanation "A+B" is always 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 analyzed how do the prior level and the causal strength level (both in terms of high or low), and their interaction, contribute to the probability of choosing one of the three options. Because the probability of choosing each option has to sum to one, which means the degree of freedom for dependent variable is $3 - 1 = 2$, we chose the explanation "A" and "A+B+C" to be the reference categories for fitting. 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 mixed-effect approach models the data in a nested structure, first level being the individual differences between participants, then the prior and causal strength condition manipulation within each participant.

Regression shows that the baseline probability of choosing "A+B+C" is higher than random (β=2.08, 95% Highest Density Interval [0.89, 3.45]). In the higher causal strength conditions the probability of choosing "A" is marginally higher (β=1.66, 95% Highest Density Interval [0.00, 3.40]). None of the other factors significantly contribute to the explanation preference. For full regression results, see Appendix Table 3.7.

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 level of causal strength difference does not significantly modulate the choice probability. Again we performed mixed-effect regression on this group of trials, all the same set up as above except the factor of causal strength is changed to causal strength difference between the causes. Only the baseline probability of choosing 1-cause is significantly higher than baseline (β=2.34, 95% Highest Density Interval [1.12, 3.69]) and none other factors are significant. Details are summarized in Appendix Table 3.7.

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:

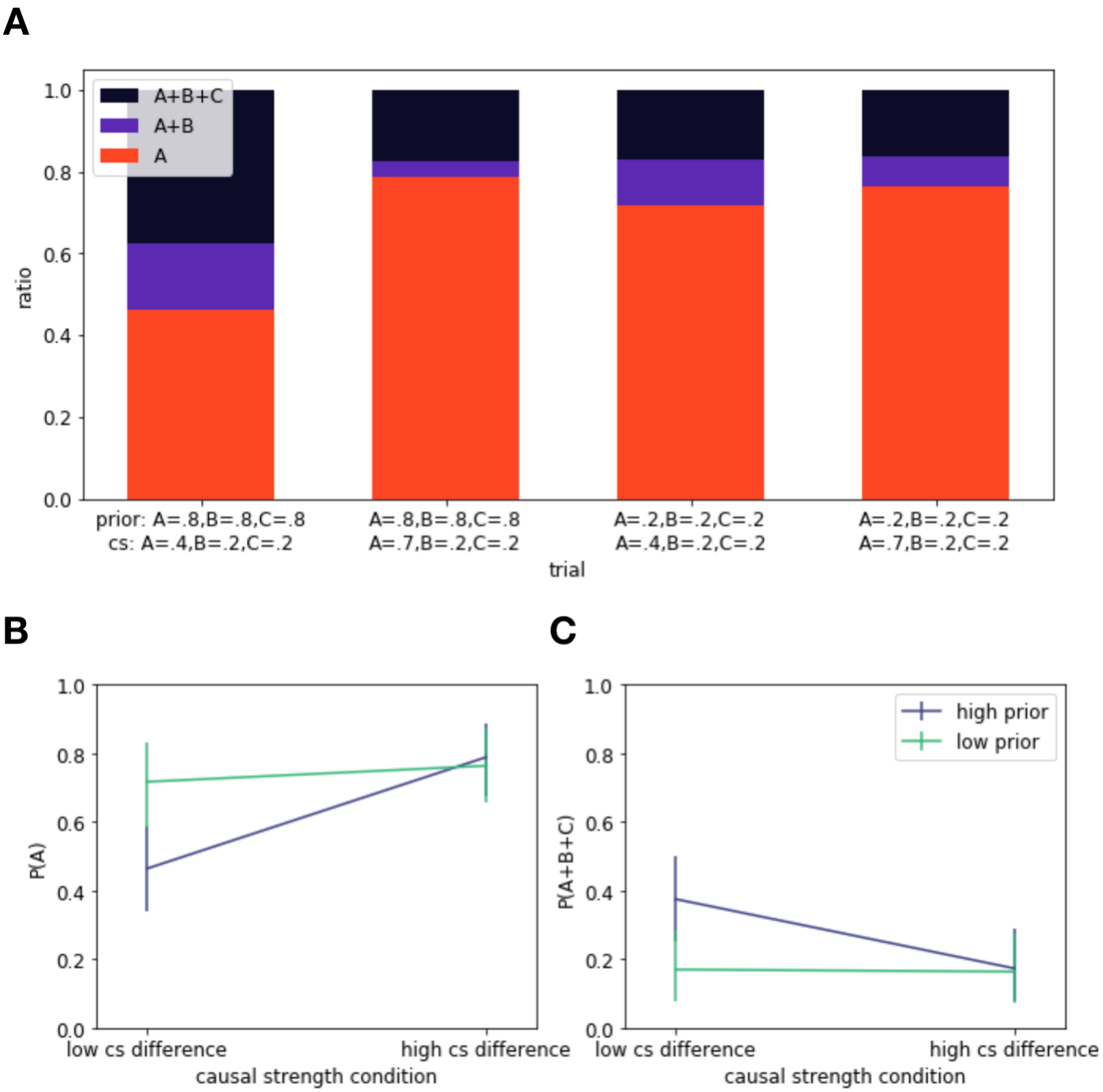


Figure 3.9: Empirical data of trials with equal causal strength and equal prior. A: Averaged ratio of choosing 1-, 2- or 3-cause; error bars indicate the 95% confidence interval calculated from bootstrapping. In the legend, "p" denotes prior and "cs" denotes causal strength of each trial. B: average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

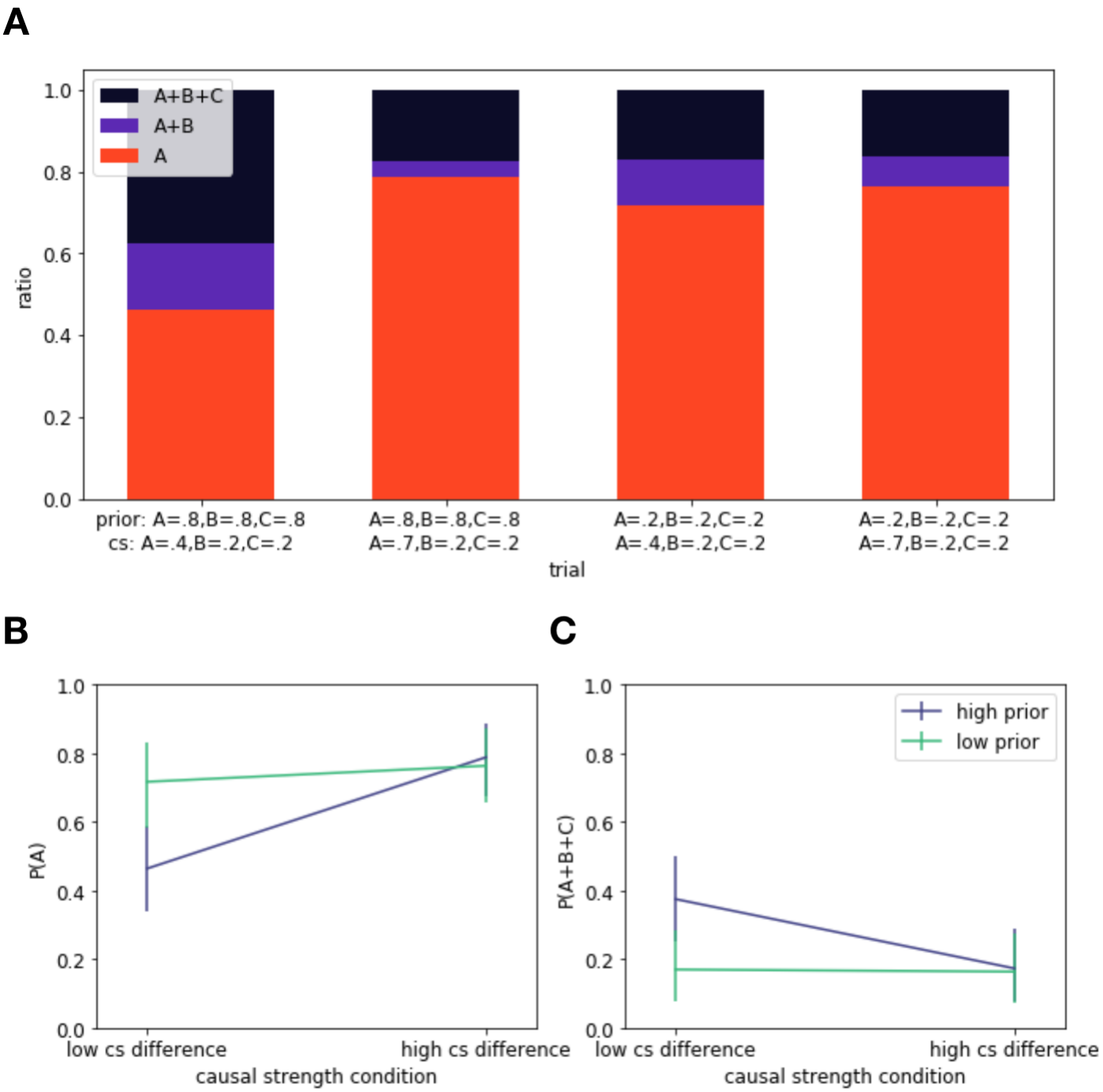


Figure 3.10: Empirical data of trials with unequal causal strength and equal prior. A: Averaged ratio of choosing different explanations; error bars indicate the 95% confidence interval calculated from bootstrapping. In the legend, "p" denotes prior and "cs" denotes causal strength of each trial. B: 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.

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 the complex explanation (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).

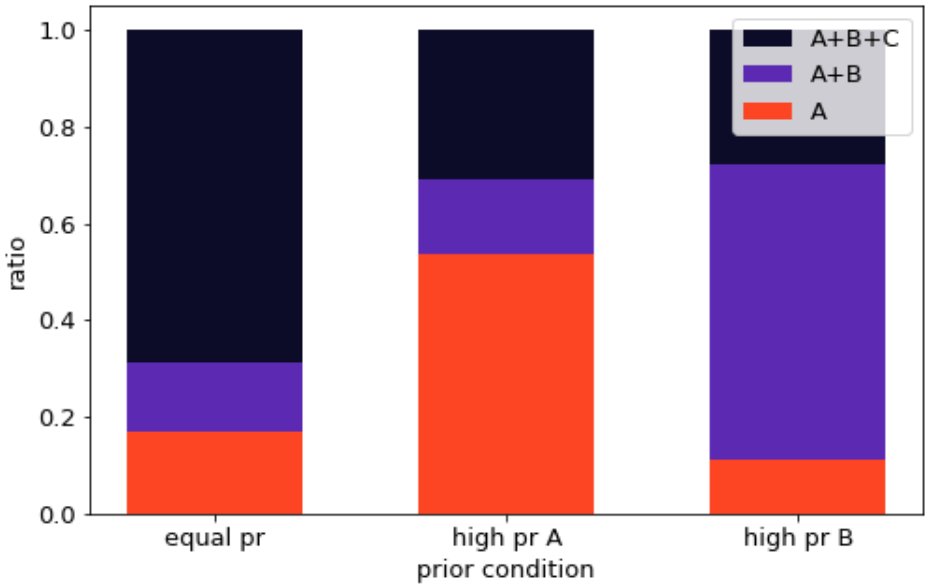


Figure 3.11: Empirical choice probability for trials of equal causal strength but equal or unequal prior. We categorized trials by their prior condition, by whether the trial has equal prior, or the first cause having the highest prior ("high pr A" in the axis label), or the second having the highest prior ("high pr B"). For each trial type, the ratio of choosing different explanations is shown in different colors.

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.

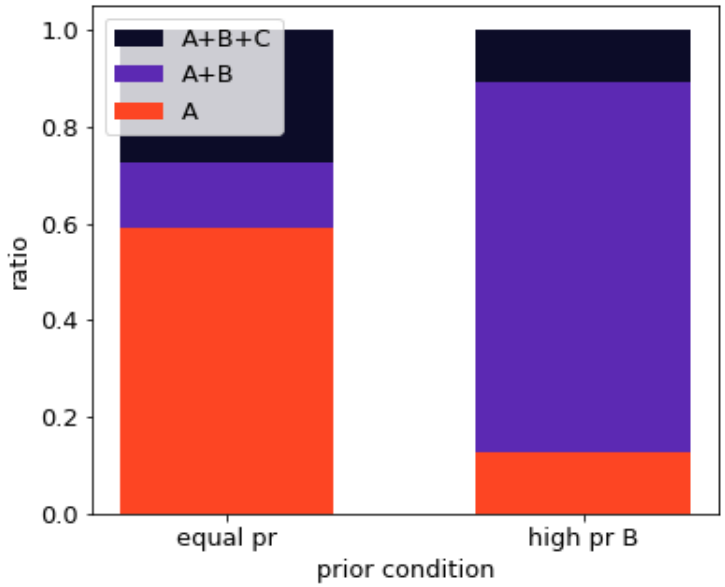


Figure 3.12: Empirical choice probability for trials of unequal causal strength but equal or unequal prior. The first cause always has the highest causal strength. We categorized trials by their prior condition, by whether the trial has equal prior, or the second having the highest prior ("high pr B").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

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 3.3 for forced choice data and Table 3.4 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 model with a softmax function that allows some randomness, so that the other explanations can be

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

Table 3.3: Comparing the predicted best explanation with empirical average best explanation. The prior and causal strength columns list the values in the order of A, B and C. 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	cs heuristic best	full heuristic 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 3.4: Comparing the predicted best explanation with empirical average best explanation for the free response data. The prior and causal strength columns list the values in the order of A, B and C. "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%.

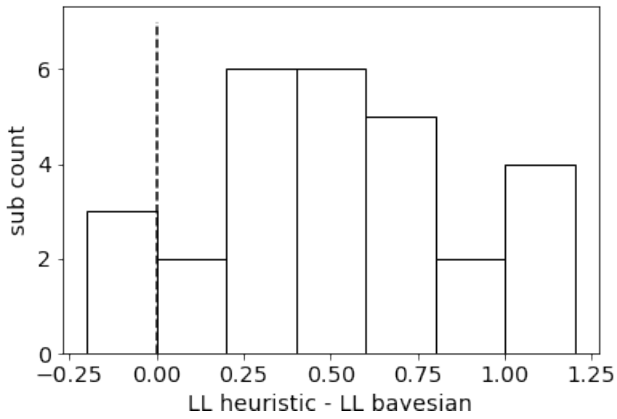


Figure 3.13: The likelihood difference between heuristic model and Bayesian model for each participant.

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.

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 3.5, 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

trial	prior	causal strength	empirical best	cs heuristic best	full heuristic 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 3.5: Comparing the predicted best explanation with empirical average best explanation for data in Experiment 1. The prior and causal strength columns list the values in the order of A, B and C. The "cs heuristic" is the one we proposed in Experiment 1 while the "full heuristic" is the one we presented in the current section.

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 3.6 and 3.7 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:

> 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.

3.8.2 Additional results

3.8.2.1 Experiment 2: regression analysis

For trials with equal prior and causal strength, the regression results are listed in Table 3.6.

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 3.6: 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.

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

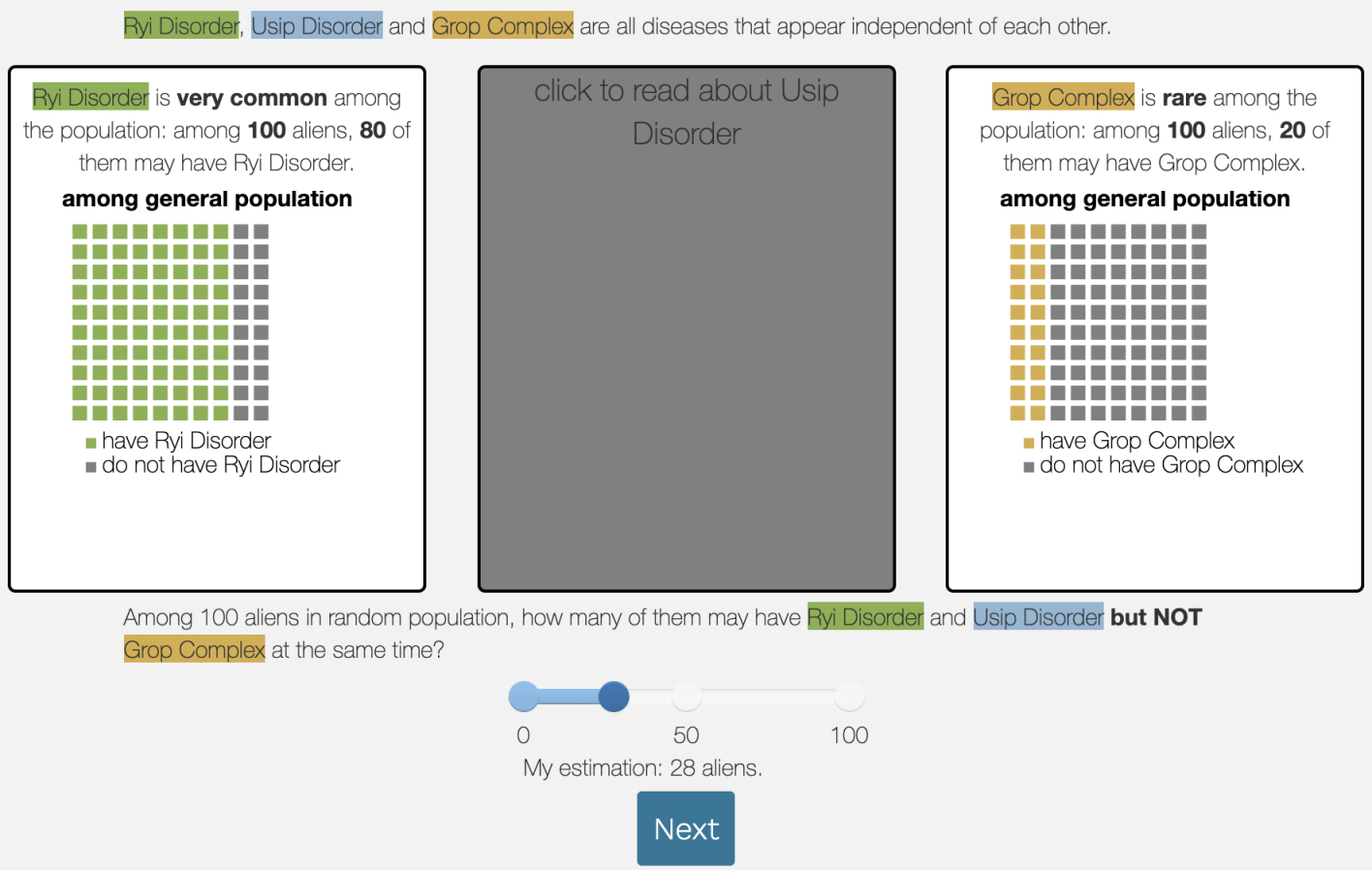


Figure 3.14: An example trial for the prior judgement section.

For trials with equal prior and distinctive causal strength for cause A, the regression results are in Table 3.7.

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.

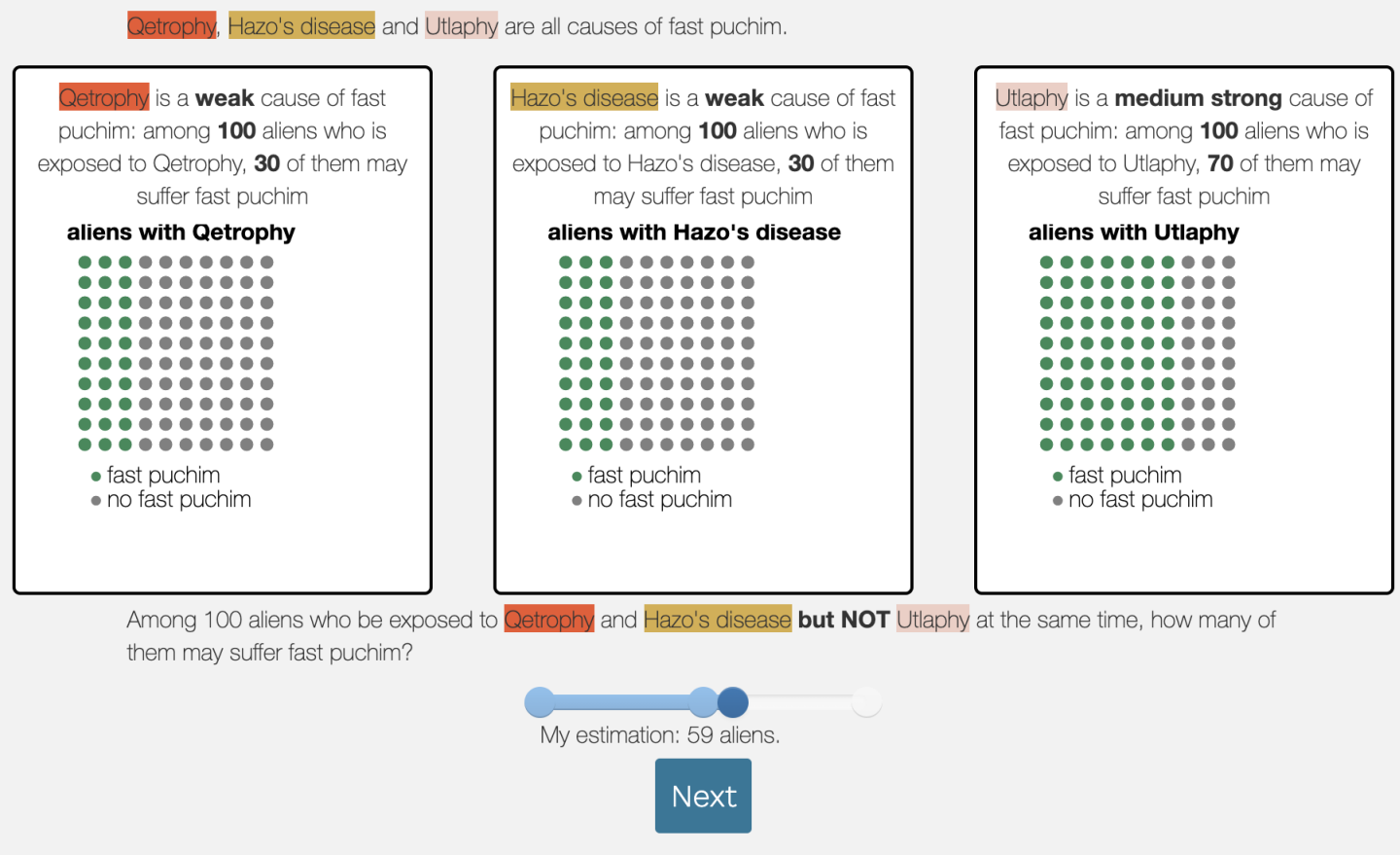


Figure 3.15: An example trial for the causal strength judgement section.

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.

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

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.

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.

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 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 3.18 is again very similar to Experiment 1 (Figure 3.8 in the main text), showing systematic deviation from the standard probabilistic theory.

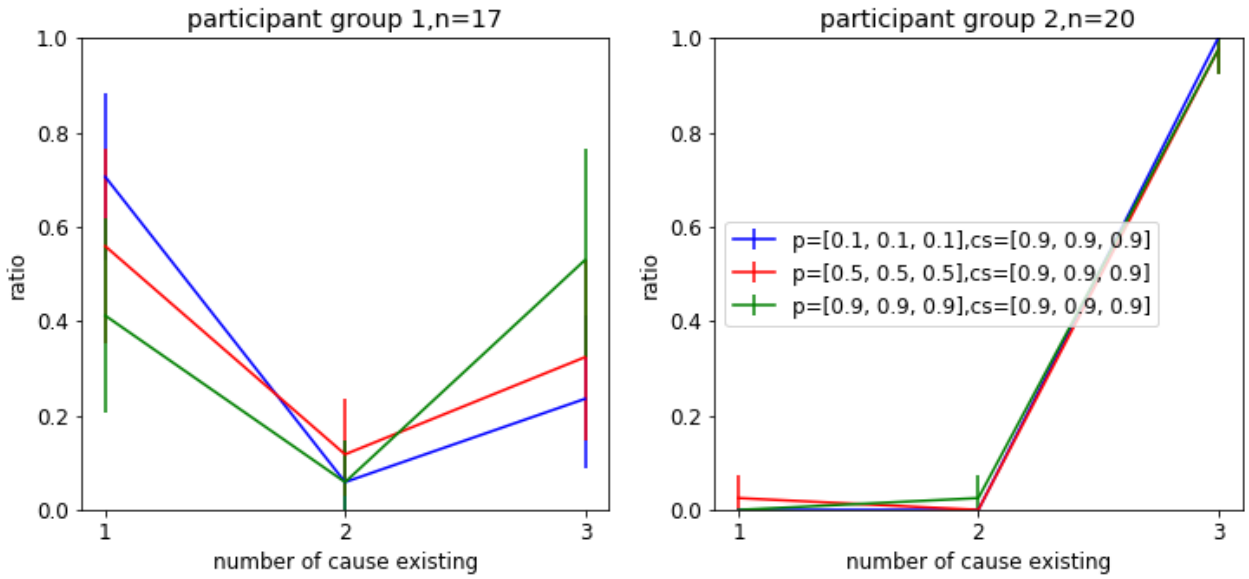


Figure 3.16: Choice ratio of the three explanations, separated by the two clusters of participants.

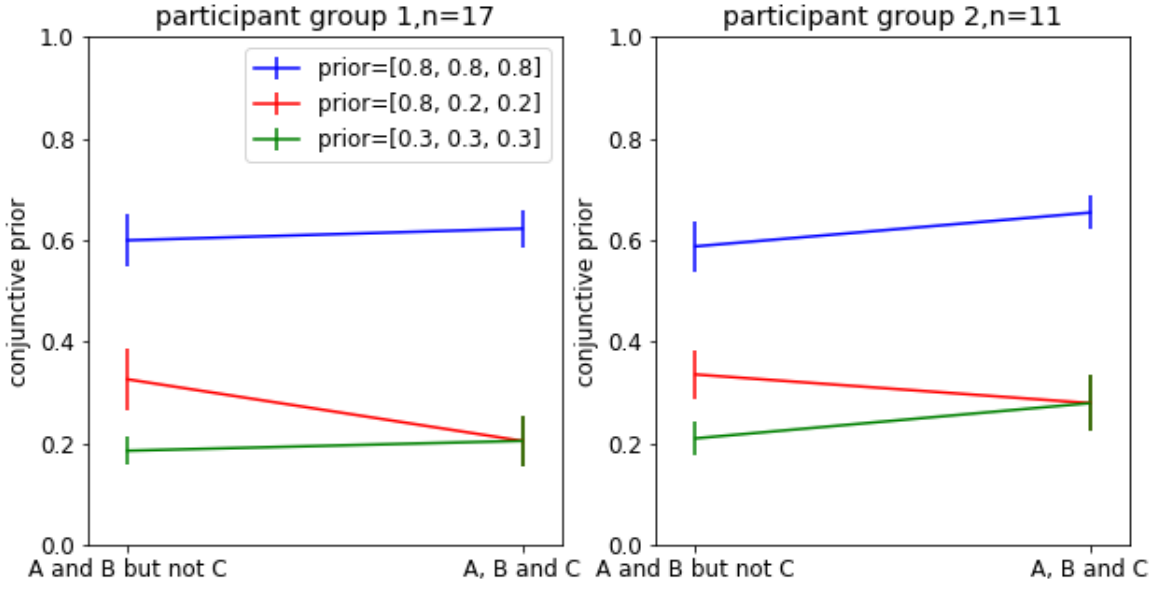


Figure 3.17: Rating of conjunctive priors, separated by the two clusters of participants, for Experiment 2.

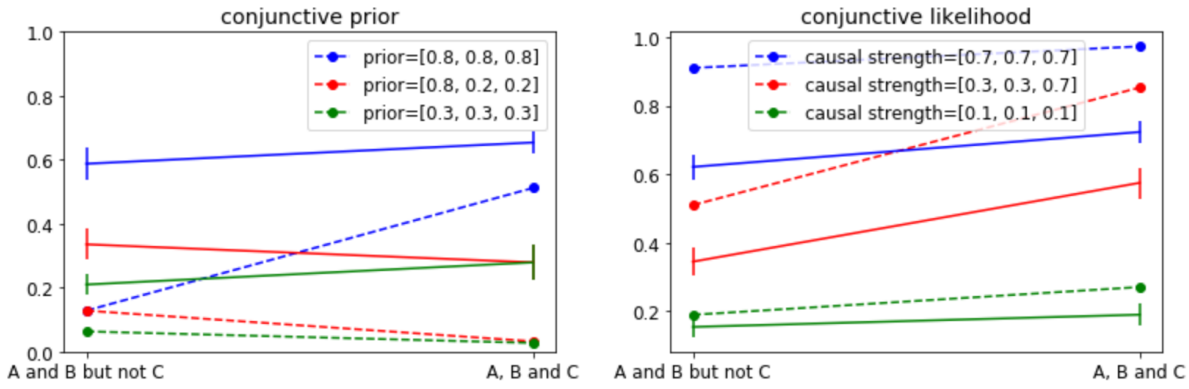


Figure 3.18: Contrasting the empirical data with theory prediction (dash line) regarding the prior and likelihood (causal strength) of conjunctive causes. The error bars on solid lines represent standard error. Note that participant reports were in the range of 0-100 and here we normalize it to the range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

CHAPTER 4

Choice preference with posterior-based account

4.1 Abstract

When people view a consumable item for a longer amount of time, they choose it more frequently; this also seems to be the direction of causality. The leading model of this effect is a drift-diffusion model with a fixation-based attentional bias. Here, we propose an explicitly Bayesian account for the same data. This account is based on the notion that the brain builds a posterior belief over the value of an item in the same way it would over a sensory variable. As the agent gathers evidence about the item from sensory observations and from retrieved memories, the posterior distribution narrows. We further postulate that the utility of an item is a weighted sum of the posterior mean and the negative posterior standard deviation, with the latter accounting for risk aversion. Fixating for longer can increase or decrease the posterior mean, but will inevitably lower the posterior standard deviation. This model fits the data better than the original attentional drift-diffusion model but worse than a variant with a collapsing bound. We discuss the often overlooked technical challenges in fitting models simultaneously to choice and response time data in the absence of an analytical expression. Our results hopefully contribute to emerging accounts of valuation as an inference process. [1]

4.2 Introduction

Eye fixations affect choices in value-based decision-making. This was originally demonstrated in a consumer decision-making task by Krajbich and colleagues (Krajbich *et al.*, 2010). Their experiment consisted of two phases. In the rating phase, subjects rated 70 snack food items on a scale from -10 to 10. In the choice phase, subjects chose between two previously rated items. The authors found that subjects more often chose the item that they fixated on for longer, even

[1]The code and analysis that support the findings of this study are openly available in the Open Science Framework (OSF) at https://doi.org/10.17605/OSF.IO/8YAC9

when the subjective ratings of the two items were equal. There is also evidence that the fixations directly cause the choice bias, instead of an underlying preference causing both (Shimojo *et al.* 2003; Armel *et al.* 2008; also see Pärnamets *et al.* 2015 for moral decisions and Tavares *et al.* 2017 for perceptual decisions; Krajbich 2019 for a review on causality between attention and choice).

To quantitatively describe the decision-making process, Krajbich and colleagues introduced the attentional drift-diffusion model (aDDM). This model predicts choices and reaction times based on the subjective ratings of the two items and the sequence of fixation times. Like the traditional DDM (Ratcliff & McKoon, 2008), the aDDM assumes a hypothetical decision variable that drifts noisily towards a bound corresponding to the item with higher subjective rating. Here, however, the drift is accelerated for the fixated item. A choice is made when the decision variable reaches one of two bounds (one for each item). The aDDM accounts for choice and response time data not only in binary choice (Krajbich *et al.*, 2010), but also in ternary choice (Krajbich & Rangel, 2011) and in purchasing choices (Krajbich *et al.*, 2012).

The aDDM does not express choice behavior as the result of the maximization of a utility function. As such, it is somewhat disjoint from many other models in behavioral economics. To bridge this gap, we build on recent work in psychology that has formulated utility functions in value-based decision-making in terms of subjective beliefs (Tajima *et al.*, 2016; Song *et al.*, 2019; Polania *et al.*, 2019). In computer science, this notion is already much older, appearing for example in Bayesian Q-learning (Dearden *et al.*, 1998). In these models, the agent maintains a continuum of hypotheses about value (or item attractiveness) and computes a Bayesian posterior distribution over value, which reflects the degree of belief in each value on the continuum; the studies differ in how the posterior is subsequently used. Intriguingly, dopamine neurons also seem to encode a distribution over future rewards (Dabney *et al.*, 2020).

These previous studies introduced the notion of probabilistic inference of value, but did not compare the resulting models to the aDDM in terms of their ability to fit data. Here, we propose a new value inference model that uses the posterior distribution as the basis of a utility function with an "uncertainty aversion" component. We fitted this model to the joint choice and response time data from Krajbich *et al.* 2010, and formally compared our model against the aDDM. To anticipate our results: we cannot confidently conclude that our model fits better than aDDM, but we show that it is at least competitive, while arguably being more principled. We also point out overlooked technical intricacies in fitting both models, which might be of interest for future studies.

4.3 Models

4.3.1 Posterior-Utility-Choice (PUC) model

4.3.1.1 Background: posterior uncertainty about value

We conceptualize the value of an item as the amount of future satisfaction resulting from consuming the item. An agent typically has incomplete or imperfect information about future satisfaction. For example, when choosing a food item, the saltiness, fat content, texture, etc. of the item are only approximately known, and so is the extent to which these properties will match the agent's personal preferences.

In the perception literature, incomplete and imperfect knowledge is most often modeled using the framework of Bayesian inference (de Laplace, 1820; Knill & Richards, 1996; Pouget *et al.*, 2013; Ma & Jazayeri, 2014). In this framework, the observer uses the sensory observations to build a subjective belief about a state of the world, as captured by a posterior probability distribution $p(\text{world state}|\text{observations})$. Recent work has proposed that value, despite being highly subjective and individualistic, may be inferred in much the same way as a world state in perception (Tajima *et al.*, 2016; Song *et al.*, 2019; Polania *et al.*, 2019); we use this notion as the basis of our model. In the Krajbich task, the agent can try to infer value from physical cues, such as product brand and information on the packaging, as well as from memories of previous experiences with the category (Shadlen & Shohamy, 2016). Combining sensory and memory cues with knowledge of one's internal state will allow the agent to form a belief about the value of the item. This belief can be expressed as a posterior probability distribution $p(\text{value}|\text{cues})$.

Based on this posterior distribution, posterior uncertainty can be defined as a summary statistic, for example as the standard deviation. Intuitively, when information is accumulated for a longer time, posterior uncertainty should generally decrease. Posterior uncertainty is different from classic risk, such as one when an agent chooses between two lotteries, e.g. \$5 for sure, or a 50% chance of receiving \$10 (Harrison & Elisabet Rutström, 2008). In such an experiment, the outcome is uncertain solely due to the stochastic step that follows the agent's choice. By contrast, posterior uncertainty can exist in a fully deterministic world. The distinction between posterior uncertainty and risk is an instance of the distinction between epistemic and aleatoric uncertainty, respectively (Chowdhary & Dupuis, 2013). We illustrate this distinction and how we view it in the full decision-making process in Fig 4.1.

4.3.1.2 Model overview

The basic premises of our model are as follows: a) fixating on an item will reduce posterior uncertainty about future satisfaction, either through the acquisition of visual information, or by trig-

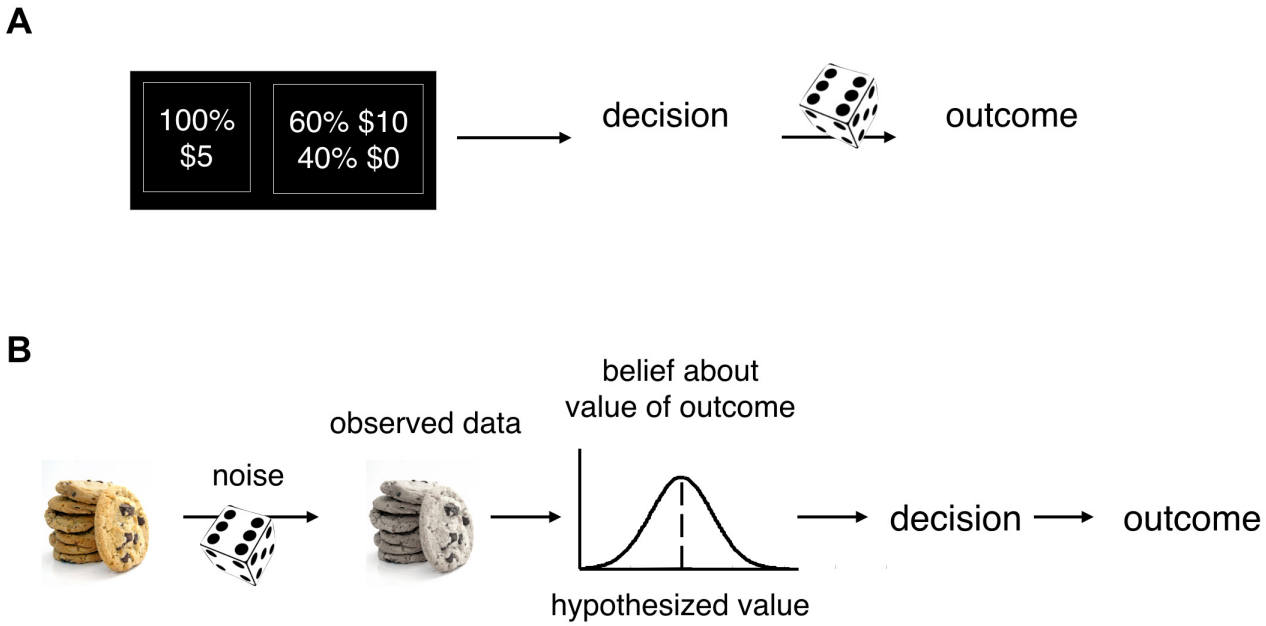


Figure 4.1: Types of uncertainty in risky choice. (A) In lotteries, uncertainty arises due to stochasticity in the mapping from decision to outcome. This is a form of aleatoric uncertainty. (B) Uncertainty can also arise from information about a choice item being incomplete or imperfect. In a probabilistic framework, the result of inferring the unknown value of a future outcome is captured by a posterior distribution, the width of which is a measure of uncertainty. This is a form of epistemic uncertainty.

gering memory recall; b) lower uncertainty leads to higher utility in an uncertainty-averse agent. By combining these mechanisms, the model can in principle account for the effect of fixation on choice: longer fixation on an item leads to lower uncertainty, which leads to higher utility, which leads to a higher probability of choosing the item. We will now turn these premises into a concrete mathematical model, comment on the relations with other models, fit the model to Krajbich's data, and compare the fit with the fit of the aDDM.

We first specify the model mathematically. In the model, when viewing a food item, the agent is trying to predict the value (future satisfaction) derived from consuming the item. On a given trial, the decision process consists of three steps:

1. the computation of a **posterior** distribution over value (with associated posterior uncertainty);
2. converting a given posterior distribution over value into a scalar **utility**;
3. converting utility to **choice**.

Accordingly, we will refer to the model as the Posterior-Utility-Choice (PUC) model. The PUC model is partially normative because it computes a Bayesian posterior in Step 1 and a utility function in Step 2. However, the form of the utility function is ad-hoc, and unlike in recent work by Tajima et al. (Tajima *et al.*, 2016), Step 3 is not normative.

4.3.1.3 Generative model

Before we can describe the decision model, we need to specify the generative model, which describes the statistical origins of the agent's data. In PUC, step 1 (posterior computation) is directly based on the generative model.

We denote the state resulting from consuming a food item by s. This high-dimensional state could comprise many factors, including satiety, nutritional state, and health status. The agent, however, does not have direct access to s, but has to form a beliefs about s from the available data D, which include observations of the properties of the item (size, ingredients, brand, etc.), interoceptive data about one's own homeostatic state (Craig, 2003), and memories of consuming similar items (Shadlen & Shohamy, 2016). We assume that in a limited time, only a limited amount of imperfect data can be collected. The mapping from s to D is stochastic and can be described by a probability distribution $p(D|s)$.

Next, we assume that the state s will produce – in a deterministic or stochastic fashion – a value v, representing an amount of future satisfaction. We model this mapping as a probability distribution $p(v|s)$. The graphical representation of the generative model is then $D \leftarrow s \rightarrow v$, where s generates both D and v.

4.3.1.4 Decision model

We now use the generative model to specify the agent's decision-making process. This process consists of three steps (P-U-C): from the sensory and memory data to a posterior distribution (Step 1), from a posterior distribution to utility (Step 2), and from utility to choice (Step 3). We now discuss each step.

- **Step 1: From data to posterior distribution over value**

  Given specific observed data D_{obs}, the agent entertains a range of hypotheses about value. The likelihood of hypothesized value v, denoted by $L(v; D_{\text{obs}})$, is now the probability that D_{obs} were produced by a state of value v:

$$L(v; D_{\text{obs}}) = p(D_{\text{obs}}|v) = \int p(D_{\text{obs}}|s)p(s|v)ds.$$

  Here, D_{obs} and s are both high-dimensional, whereas v is one-dimensional. We assume that the effect of viewing longer is that more data are gathered. Unfortunately, we know neither $p(D_{\text{obs}}|s)$ nor $p(s|v)$ since the researchers cannot observe the complete interoceptive state s. Therefore, we make a simplification, where we assume that the likelihood over v is Gaussian with mean x and standard deviation σ, which we will assume to be the same for all items:

$$L(v; D_{\text{obs}}) \propto \mathcal{N}\left(v; x(D_{\text{obs}}), \sigma(D_{\text{obs}})^2\right).$$

  Here, x is the maximum-likelihood estimate of v, the best guess one could make about v based on D_{obs}. By analogy with perception, we will refer to x as a *measurement*, and a consistent generative model for x would be $p(x|v) = \mathcal{N}(x; v, \sigma^2)$, where v is the true value. This simplification is largely analogous to the mapping from a neural population model to a behavioral model (Ma, 2010). As a proxy for the true value of consuming an item, we take the rating given by the subject for that item in the rating phase of the experiment.

  We next assume a Gaussian prior $p(v) = \mathcal{N}\left(v; \mu_p, \sigma_p^2\right)$. Including the prior, the posterior over v becomes

$$p(v|D_{\text{obs}}) \propto L(v; D_{\text{obs}})p(v), \tag{4.1}$$

  which we approximate by

$$p(v|x) \propto p(x|v)p(v). \tag{4.2}$$

  The accumulation of evidence is gated by fixations. Moreover, the longer the agent fixates on an item, the more measurements are made. Within a trial, the number of measurements

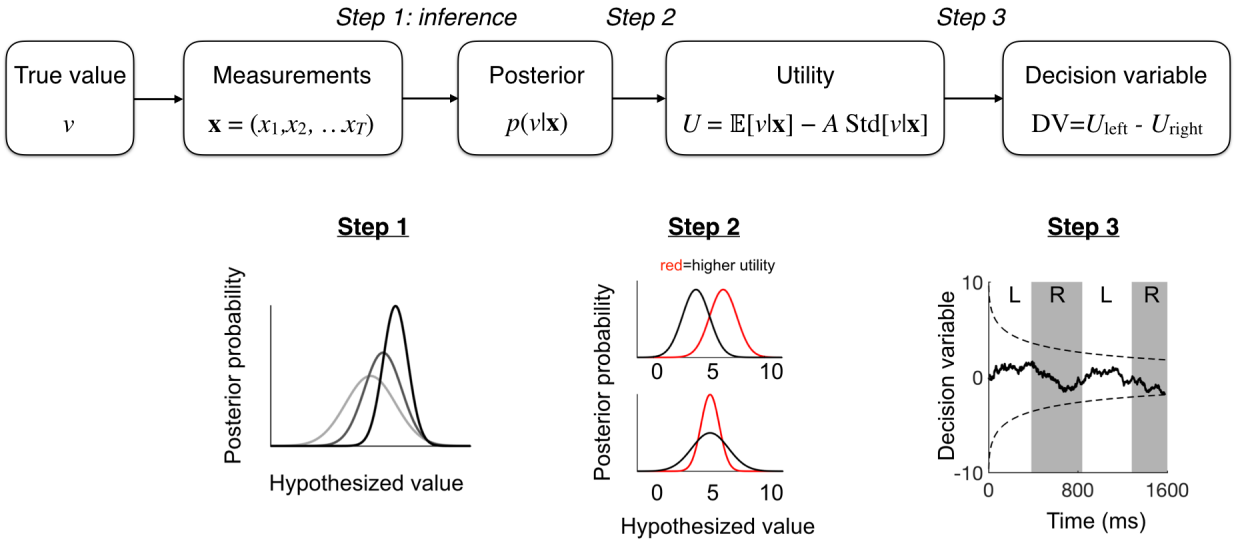


Figure 4.2: Posterior-Utility-Choice model. The PUC model describes how an agent maps noisy measurements of value to a decision variable. Top: Flow diagram of the model. Bottom: Components of the model. Step 1: The agent computes a posterior distribution over hypothesized value. As viewing time increases (darker colors), the posterior distribution shifts from the prior towards the true value, and becomes narrower. Step 2: Utility incorporates both the mean and the standard deviation of the posterior over value. Both higher mean and lower standard deviation are preferred. Utility is evaluated separately for each item. Step 3: Evolution of the decision variable on an example trial. L and R denote fixations on the left or the right item. The decision variable, DV, is the utility difference of the two items. A decision is made when DV crosses the collapsing bound (dashed).

can differ between the two items due to unequal fixation times. For each item, the posterior distribution starts as the prior distribution. As the agent gathers measurements by fixating on an item, the posterior distribution for that item updates iteratively:

$$p(v\,|x_1, \ldots, x_{T+1}) \propto p(x_{T+1}|v)p(v\,|x_1, \ldots, x_T). \tag{4.3}$$

Qualitatively, this updating has two effects: first, the posterior mean $\mathbb{E}[v\,|x_1, \ldots, x_T]$ will move away from the mean of the prior, μ_p, toward the true value, v; second, the variance of v under the posterior distribution will decrease (Fig 4.2, step 1).

- **Step 2: From posterior distribution over value to utility**

Now that the agent has a belief over value as expressed in the posterior distribution, they have to turn this belief into a utility. The simplest way is simply to take a central tendency of the posterior (mean, median, or mode of value under the posterior). But this might discard too much information. In principle, any mapping from the posterior to a number can serve to compute the utility of an item. Since the posterior is a function, utility is then a function of a function, and we will denote it by $U[p(v|\mathbf{x})]$.

Furthermore, we define utility as a weighted average of the mean and the standard deviation of value under its posterior distribution based on a sequence of measurements $\mathbf{x} = (x_1, \ldots, x_T)$ (Fig 4.2B):

$$U[p(v|\mathbf{x})] = \mu_{\text{posterior}} - A \cdot \sigma_{\text{posterior}}, \tag{4.4}$$

where A is a constant that we will call the "uncertainty aversion parameter", and

$$\mu_{\text{posterior}} = \mathbb{E}[v|\mathbf{x}] \tag{4.5}$$

$$\sigma_{\text{posterior}} = \text{SD}[v|\mathbf{x}] \tag{4.6}$$

The standard deviation term can be motivated in at least three, not mutually exclusive ways. First, it is possible that the proper definition of utility is $U = \mathbb{E}[v|D_{\text{obs}}]$, i.e. the expected value of the posterior based on D_{obs}. We made the approximation that the likelihood over v is Gaussian, but if it is not, an error will be introduced by instead using $\mathbb{E}[v|x]$. The standard deviation term could partially compensate for this error, in a way similar to Levy & Markowitz 1979. Second, even when the likelihood is exactly Gaussian, the agent might be uncertainty-averse. The uncertainty arises from incomplete and imperfect data (i.e. stochasticity in the measurements), rather than from post-decision stochasticity, as in classic risk aversion. Standard models of risky choice such as in 4.1A are a specific case of our model. In the classic risk models, the future state s is typically the state of having received a specific monetary amount, and no data need to be collected to know the distribution of s. Moreover, the mapping from s to v is typically assumed deterministic and monotonic, say $v = F(s)$. Then Eq. (4.1) for the posterior over v reduces to $p(v) = \int \delta(v - F(s))p(s)ds$ (all of this can be conditioned on an action). Thus, our model generalizes standard models of risky choice by replacing the distribution $p(v)$ by a posterior distribution $p(v|D_{\text{obs}})$. A term similar to the standard deviation term has been used in portfolio theory (Markowitz, 1952). Third, we are inspired by the literature on the so-called *mere exposure effect* (Zajonc, 2001), the phenomenon that mere exposure to a stimulus (for example a consumer item or a work of art) increases the observer's preference for this stimulus. A leading explanation of the mere

exposure effect is uncertainty reduction (Bornstein, 1989; Lee, 2001), the idea that people prefer stimuli that are familiar.

Under our Gaussian assumptions for likelihood and prior, the mean and standard deviation of the posterior become

$$\mu_{\text{posterior}} = \frac{\frac{\mu_p}{\sigma_p^2} + \frac{T\bar{x}}{\sigma^2}}{\frac{1}{\sigma_p^2} + \frac{T}{\sigma^2}} \tag{4.7}$$

$$\sigma_{\text{posterior}} = \frac{1}{\sqrt{\frac{1}{\sigma_p^2} + \frac{T}{\sigma^2}}}, \tag{4.8}$$

where $\bar{x}$ is the mean of the measurements $(x_1, \ldots, x_T)$. Increasing viewing time will change utility in two ways: by moving the posterior mean from μ_p to the true value v, and by decreasing the posterior standard deviation. The former of these changes can be negative or positive depending on the prior estimation of value and the specific value of the current item, the latter is always positive. In Fig 4.2, step 2 we show two examples of how one item can have higher utility than the other.

- **Step 3: From utility to choice**

  The final step is to map utility to choice. We posit that the agent's decision variable, denoted by DV, is the difference between the utilities of the two items:

$$\text{DV} = U_{\text{left}} - U_{\text{right}}. \tag{4.9}$$

  Finally, the agent terminates the decision process when DV crosses a decision bound. The agent will then choose the item with the higher utility (Fig 4.2, step 3). We choose the decision bound to be decreasing over time ("collapsing"). A general motivation for using a collapsing bound instead of a fixed one is to prevent the model from predicting unrealistically long reaction times when deciding between two very similar items (Churchland *et al.*, 2008; Hawkins *et al.*, 2015; Milosavljevic *et al.*, 2010; Drugowitsch *et al.*, 2012). For the specific form of collapsing bound, we use a special case of the Weibull function suggested by (Hawkins *et al.*, 2015):

$$B_t = B_0\, e^{-\left(\frac{t}{\lambda}\right)^k}. \tag{4.10}$$

  An example of the evolution of the decision variable and the decision bound is shown in Fig 4.2, step 3).

  In addition, we allow for non-decision (or residual) time τ. This means that the decision process may end before the fixation series has ended, and the remaining time would then not

contribute to the decision, regardless of how many fixations occur in that time; in practice, the estimated non-decision time is so short that fixation rarely switches in this time. Our non-decision time is conceptually different from the one in Krajbich & Rangel (2011) and Krajbich *et al.* (2012): theirs can be interpreted as accounting for the transitions between fixations, while ours is part of the total fixation time and interpreted as the time spent looking at items after the decision has been made.

Finally, we add a guessing rate parameter g. For each trial, there is a probability g of making a random decision at a random moment (based on the empirical decision time distribution fitted by a Weibull function, details see Appendix 4.7.2), then the rest probability of $1 - l$ making decision based on the model prediction. This is necessary to avoid a zero probability for a smooth likelihood function landscape that allows parameter fitting.

4.3.2 Attentional drift-diffusion model

Krajbich and colleagues (Krajbich *et al.*, 2010; Krajbich & Rangel, 2011; Krajbich *et al.*, 2012; Krajbich & Smith, 2015) have proposed the attentional drift diffusion model (aDDM), which conceptualizes the consequences of fixation as biasing the drift velocity of the fixated item. The decision variable is defined as:

$$\mathrm{DV}_t = \mathrm{DV}_{t-1} + d(r_{\mathrm{left}} - \theta r_{\mathrm{right}}) + \epsilon_t \tag{4.11}$$

when fixating on the left item, and

$$\mathrm{DV}_t = \mathrm{DV}_{t-1} + d(\theta r_{\mathrm{left}} - r_{\mathrm{right}}) + \epsilon_t \tag{4.12}$$

when fixating on the right item, where d is a scaling constant, r_{left} and r_{right} are the ratings for the two items, and ϵ_t is diffusion noise, drawn independently across time points from a normal distribution $\mathcal{N}(0, \sigma^2)$. aDDM differs from the standard drift-diffusion model (Ratcliff & McKoon, 2008) in the attentional bias factor θ, which takes values between 0 and 1. Diffusion continues until the DV hits one of two boundaries, which are assumed symmetric with respect to 0. In the original work by Krajbich and colleagues (Krajbich *et al.*, 2010), the boundaries were fixed over time. As usual, there is an arbitrary scaling in aDDM which allows us to set $B_0 = 1$; in PUC, this is not possible because the scale is already set by v, which we approximate by the subject's rating of the item.

aDDM with collapsing bounds. We also consider a more flexible variant of the aDDM, namely one that has the same parametric family of collapsing bounds as the PUC model (Eq. (4.10)). We call the resulting model the attentional collapsing-bound drift-diffusion model (acb-

DDM), by analogy to the non-attentional version, which has been called cbDDM, with "cb" standing for "collapsing bound". Milosavljevic *et al.* (2010) previously considered a collapsing bound that is a special case of ours, namely with $k = 1$. In principle, we could also allow for an increasing bound by changing the parameter ranges of the boundary function. We limit ourselves to a collapsing bound here, following previous work and because it is psychologically easier to interpret.

Differences between PUC model and aDDM. The PUC Model and the a(cb)DDM have mechanistic similarities: in both models, the agent decides when the decision variable crosses a bound, allowing the model to make predictions for the relation between choice and total fixation time. However, the PUC model differs conceptually from the a(cb)DDM in the following aspects:

(a) the PUC agent chooses the item with the highest utility, whereas the a(cb)DDM does not have an immediate interpretation in terms of utility. (This stands in contrast to the basic DDM model, in which the decision variable can be interpreted as the difference between the values of the two items (Webb, 2019));

(b) in the PUC, noise is specifically associated with the agent's observed sensory information and retrieved memories, whereas the origin of noise in the a(cb)DDM is not well specified;

(c) In the PUC model, later measurements have a smaller effect on the decision variable than earlier ones, because all measurements are generated from the same distribution and there are diminishing returns to information collection as the estimated value approaches the underlying true value. By contrast, in the a(cb)DDM, the variance of the noise added at each time point stays the same across time;

(d) in the PUC model, two main mechanisms influence the preference: the uncertainty reduction term is independent of the item rating, indicating an additive effect of attention, whereas the posterior mean update depends on the difference between the prior mean value and the specific item value, thus indicating a multiplicative effect of attention. In contrast, a(cb)DDM will always boost the item with higher value more, thus the influence of attention is always multiplicative. For more studies regarding the additive versus multiplicative nature of attention, see Cavanagh *et al.* (2014); Smith & Krajbich (2019); Westbrook *et al.* (2020).

In addition, on the surface, it seems that in the PUC model, the agent keeps two distinct value distributions, one for each item, whereas in the a(cb)DDM, the information of the two items are combined into a single "relative decision variable". However, in the later extension of aDDM model where more than two options are being compared (Krajbich & Rangel, 2011), separate accumulators for each item were used and and it could be reduced to the aDDM with two items. Thus, the a(cb)DDM can also be conceptualized as two distinct accumulators, followed by another step of deriving the relative value difference, which is similar to the PUC model.

4.4 Materials and Methods

Data. The data from the experiment in Krajbich *et al.* (2010) were made available to us by the authors. The data set contained 39 participants with an average of 95 trials per participant (25 participants completed the maximum number of 100 trials). On each trial, the data of interest consisted of the previously collected ratings of the presented items and the eye fixation series summarized as a binary sequence with values "left" and "right", with the corresponding fixation times. We are not fitting reaction times, but instead total fixation time.

Likelihood. The PUC model, as introduced in "Decision models" above, has 4 parameters for the value estimation (σ, σ_p, μ_p,A), three bound parameters (B_0, k, and λ), one guessing rate parameter (g) and a non-decision time parameter τ. To simplify, we fixed the prior mean σ_p and variance μ_p to be the empirical mean and variance extracted from the rating data, thus leaving 5 parameters to be fitted. We tested the more flexible versions of PUC too, as well as another reduced PUC model(see summary in the "Result" section and details in Appendix 4.7.1). The aDDM has 3 parameters for the drift process: σ, d, θ, two bound parameters (k and λ, since in their model, B_0 can be set to 1 without loss of generality). To match with PUC for a fair model comparison, we added a guessing rate parameter (g) and a non-decision time parameter τ for aDDM.

We fitted the parameters in each model on an individual-subject basis using maximum-likelihood estimation. The inputs to the model on the ith trial consist of the ratings of the left and right items, $r_{\text{left},i}$ and $r_{\text{right},i}$. The model predicts the joint probability of the choice C_i and the total fixation time T_i. The log likelihood of the parameters is then

$$\log L(\text{parameters}) = \sum_{i=1}^{n_{\text{trials}}} \log p(C_i, T_i | r_{\text{left},i}, r_{\text{left},i}, \text{fixation series}_i, \text{parameters}). \tag{4.13}$$

We did not constrain T_i to be later than all of the observed fixations. Thus, the model is allowed to wrongly predict a total fixation time that falls somewhere in the middle of the fixation series.

Fitting procedure. We used maximum-likelihood estimation to fit the parameters, separately for each participant. This involves maximizing Eq. (4.13), which in turn involves calculating for a given parameter combination and on each trial the probability that the model observer produces the participant's choice C_i with the participants' total fixation time T_i. To calculate this probability, we numerically propagated the probability distribution of the decision variable (See Eq. 4.9 and Eq. 4.11-4.12) across time. At each time step, we used the boundaries to truncate the distribution, with the truncated probability being our estimate of the probability of the corresponding response. The (non-normalized) remaining part of the distribution is propagated further.

To maximize the log likelihood, we used Bayesian Adaptive Direct Search (**?**), a global optimization algorithm that uses Bayesian methods to approximate the shape of the likelihood function. To minimize the risk of getting stuck in local optima, we ran the optimization algorithm with multiple initial conditions (see Appendix 4.7.2).

Differences from Krajbich et al. (2010). Our fitting procedure differs from the one used in the original paper (Krajbich *et al.*, 2010) in the following aspects:

- Instead of fitting parameters at the group level, we fitted parameters to individual subjects. This is more accurate if individuals differ from each other.

- Krajbich et al. binned the data in 100 ms bins, but simulated the time course of their model evolution in 1 ms steps. Instead, we used steps of 100 ms for the latter as well. Apart from saving computational time, this choice is more consistent with the assumption that measurements are independent across time points (in view of neural autocorrelation functions). Moreover, 100 ms is still reasonably fine compared to the median total fixation time (which was about 1.4 seconds; mean being 1.9 seconds).

- Instead of fitting the reaction time, we fitted total fixation time, in an attempt to avoid epochs in which the observer did not fixate on either item.

- Krajbich et al. used randomly sampled fixation durations from the empirical distribution to perform the simulation both in this and later work (Krajbich & Rangel, 2011; Krajbich *et al.*, 2012). Instead, we used the actual fixation data for each trial, simulating only until the time when the actual fixation has ended and calculating the probability that the choice was made at the end of the empirical fixation series. All remaining probability went into a single bin representing later decision times. Note one exception is when we plot the summary statistics of probability of making choice against the total fixation time. We used the distribution of fixation times split out by subject, then independently and sequentially drew from this distribution to create a synthetic fixation series for each trial after the empirical fixation series has ended; we repeated this 10 times for each trial which is enough to achieve a stable result. This allows us to obtain an unrestricted model prediction for total fixation times (which we use as a proxy for reaction times throughout the paper).

- Krajbich et al. performed 1000 simulations to calculate the log likelihood for each rating pair (3000 in their later work, see Krajbich *et al.* 2012). However, this number is low relative to the number of possible responses (in Krajbich et al., 2 choices times 52 reaction time bins; for us, 2 choices times 50 total fixation time quantiles). In such situations, estimating the log likelihood through brute-force simulation not only causes variance to be high, but is

also biased due to the nonlinearity of the logarithm (van Opheusden *et al.*, 2020). This problem is particularly stark when the simulation assigns zero samples to an observed response. Therefore, instead, we used the numerical approximation mentioned above.

- Instead of using a grid search in parameter space, we used Bayesian Adaptive Direct Search (Acerbi & Ma, 2017), which is a more precise and more reliable optimization method.

Parameter recovery and model checking. To confirm the validity of our model fitting choices, we fitted synthetic data using the same fitting procedures as for the real data. To generate synthetic data, we used the exact same rating distribution as the real data by matching each synthetic trial with a real trial. Each synthetic subject was given the parameters that best fitted one real subject. Then we performed fitting for individual synthetic subjects using methods introduced above. Results are presented in Appendix 4.7.3. The summary statistics are recovered very well. Parameter recovery is generally good but somewhat worse for the more complex models (acbDDM and PUC). This is likely due to soft trade-offs between parameters. As a result, the parameter estimates in the real data should be taken with a grain of salt. However, the results of our paper do not rely on parameter estimates but only on log likelihoods and summary statistics. Therefore, the results are not affected by issues with parameter recovery.

Model comparison. To compare models, we used the corrected Akaike Information Criterion (AICc; Akaike 1974; Burnham & Anderson 2004) and the Bayesian Information Criterion (BIC; Schwarz *et al.* 1978).

4.5 Results

To model the effects of fixation on choice, we introduced the Posterior-Utility-Choice (PUC) model, in which the agent judges the value of an item by (1) accumulating evidence, gated by fixations; (2) computing and updating two posterior probability distributions over value, one for each item; (3) calculating the utility of each item not only from its posterior mean but also from its posterior standard deviation, with the latter accounting for uncertainty aversion. We compared the PUC model to the established attentional Drift Diffusion Model (aDDM). We fitted model parameters to individual-subject data using maximum-likelihood estimation.

PUC versus aDDM. To compare the goodness of fit of the PUC model with that of the aDDM, we first inspected model fits to several summary statistics plotted by (Krajbich *et al.*, 2010) (Fig 4.3): the proportion of choices of an item as a function of the fixation time advantage for that item, the same but conditioned on the item's rating, and the proportion of choices of an item as a function of the rating difference between the two items and which item was fixated last. To obtain

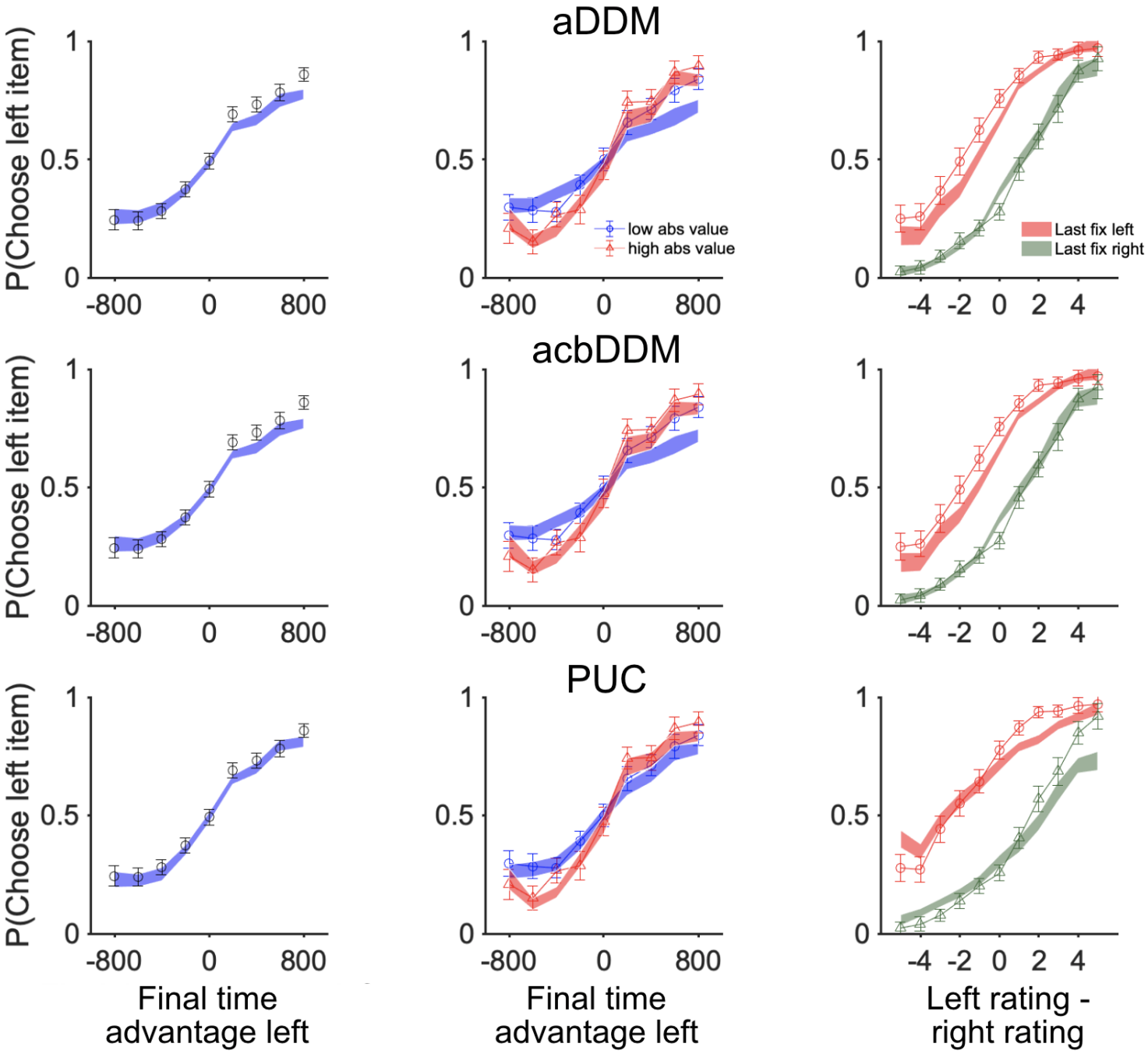


Figure 4.3: Fits of the aDDM, the acbDDM, and the PUC model to summary statistics of the data. (A) When the total fixation time advantage of an item increases, that item is chosen more often. (B) Same as A but conditioned on item rating. Both models predicted that when the absolute values of both items are higher (i.e. both are more preferred items), the fixation modulation effect is larger, which trend is less significant in the empirical data. (C) Besides total fixation duration, the last fixation also biases the choice. (D) Distribution of total fixation time. The aDDM fits we obtained differ from those in Krajbich *et al.* (2010) not only because of differences in parameter estimation methods (see Method section 4.4, *Differences from Krajbich et al. (2010)*), but also because of a difference in trial aggregation (see below - *PUC versus aDDM*).

the fits of a model to a summary statistic, we ran each model in generative mode using the fitted parameters for an individual subject to obtain a model prediction for each trial, then aggregated these predictions to compute the summary statistic.

Both the PUC model and the aDDM qualitatively capture the behavioral phenomena that an item viewed for a longer time (Fig 4.3A-B) or viewed last (Fig 4.3C) is more likely to be chosen. However, both models show modest deviations from the data. In addition to choice data, we also plotted the distribution of the total fixation time (Fig 4.3D), which shows a somewhat worse fit for the aDDM than for the PUC model.

Note that our fits of the aDDM to the summary statistics look different from those given by Krajbich *et al.* (2010). This is mainly because of a difference in trial aggregation. Krajbich et al. aggregated the model predictions across rating pairs without taking into account the frequencies of these pairs in the experiment. We instead aggregated individual-trial predictions; this is necessary in our case because we used the individual-trial fixation series, but it also ensures a proportional representation of each rating pair in the summary statistics.

To quantitatively compare the models, we performed formal model comparison (Table 4.1). The PUC model has a lower AICc than the aDDM (thus a better fit), with a summed difference across subjects of -461. The bootstrapped 95% confidence interval was $(-887, -136)$. BIC penalizes the number of parameters more, causing the 95% confidence interval to include 0. Overall, the PUC model performs better than the aDDM.

Collapsing bound. Next, we included the aDDM with collapsing bound, acbDDM, into the comparison. The fits of the acbDDM to the summary statistics are similar to those of the aDDM (Fig 4.3), but the acbDDM performs substantially better than the aDDM in model comparison (summed AICc difference: -714; summed BIC difference: -541). The acbDDM model also clearly outperforms the PUC model (summed AICc and BIC difference: 253; these two metrics are same because both models have the same number of parameters).

Hierarchical Bayesian model selection. So far, we assumed that every subject followed the same model. However, there might be heterogeneity in the population. To account for this possibility, we performed hierarchical Bayesian model selection (Stephan *et al.*, 2009) using the VBA package (Daunizeau *et al.*, 2014). According to this analysis, the proportions of the population following the PUC, the aDDM, and the acbDDM are 20.5%, 25.6%, and 53.9% based on AICc (and 18.0%, 35.9%, and 46.1% based on BIC). However, more work would be needed to conclusively establish heterogeneity in the population.

PUC model variants. Finally, we examined three variants of the PUC model: first, a variant in which the prior variance is a free parameter instead of a fixed constant; second, a variant in which both the prior variance and the prior mean are free parameters; third, a variant without the uncertainty term in Eq. (4.4) (i.e. $A = 0$). We found that all three variants fit worse than the main

	PUC-aDDM	PUC-acbDDM	acbDDM-aDDM
neg LL	-321 (-533, -160)	127 (55, 197)	-448 (-613, -318)
AICc	-461 (-887, -136)	253 (110, 394)	-714 (-1045, -456)
BIC	-288 (-708, 35)	253 (110, 394)	-541 (-869, -284)

Table 4.1: Comparing the main PUC model to alternative models according to negative log likelihood (not corrected for the number of free parameters), AICc, and BIC. Lower values are better for the first-mentioned model. All values are summed across subjects; bootstrapped 95% confidence intervals are given in parentheses.

model we presented in the paper, in terms of both AICc and BIC (see Appendix 4.7.1).

4.6 Discussion

In this work, we use the recent idea that valuation is a form of Bayesian inference to explain the effect of fixation on choice. In the Posterior-Utility-Choice (PUC) model, the agent continuously updates a posterior distribution over the value of an item based on a sequence of noisy measurements and computes the utility of that item as a weighted difference between the posterior mean and posterior uncertainty, the latter reflecting uncertainty aversion. The decision is made when the utility difference between the two items reaches a bound. We found that the PUC model accounts better for the effects of fixation on choice than the original model, the attentional drift-diffusion model (aDDM), but not better than its generalization with a collapsing bound, the acbDDM; thus we provided some supports for a flexible bound as a decision model component (although, evidence is mixed in the literature, see Milosavljevic *et al.* 2010; Hawkins *et al.* 2015).

Setting aside goodness of fit, the PUC model is a different type of model than the Attentional Drift-Diffusion models. The PUC model postulates what the agent cares about at a behavioral level, through a utility function derived from a posterior distribution over value. By contrast, the a(cb)DDM is neither stated in terms of utility nor involves computing a belief over value.In addition, the PUC model makes a more explicit commitment to the origin of behavioral variability: ultimately, it stems from the noise in sensory measurements or retrieved memories. The a(cb)DDM does not make such a commitment.

These conceptual differences between the PUC model and the a(cb)DDM allow for novel predictions. First, the role of the uncertainty and therefore the utility of an item can be studied more explicitly in a new experimental design. Other than explicitly measure the uncertainty in the rating phase like in Gwinn & Krajbich (2020), future experiments could even manipulate the uncertainty by compromising the quality of the data that the subject receives about the value of an item. This could be done through a simple visual manipulation, such as lowering contrast or blurring the im-

age, or through a memory manipulation, such as presenting items that differ in the time elapsed since the subject last interacted with them. We predict that lower quality of data will lead to lower utility and in turn to the item being less likely to be chosen. Second, a similar prediction could be tested in a timed rating experiment, where speeded judgments should lead to lower ratings. Third, we predict that changing the prior distribution in Eq. (4.2) will affect choices in a specific manner. For example, consider a choice between two items with a true value of 3, yet the participant has a prior estimate of 0. Fixating longer on one item then has two effects on the utility of that item: it increases the posterior mean (since the likelihood mean is greater than the prior) and it decreases posterior uncertainty. Now consider the same agent but with a prior mean of 6. Then, longer fixation has two counteracting effects on utility: it will still decrease uncertainty, but now it will *decrease* the posterior mean. Comparing to the prior mean of 0, we expect a weaker or even reversed effect of fixation on choice. Thus, it might be interesting to experimentally manipulate the prior distribution.

Our work has several limitations:

- All extant models exhibit noticeable deviations from the data, which leaves a challenge for future modelers.

- We assumed a specific direction of causality: that fixation increases preference, instead of the other way round. Both the PUC model and the aDDM assume this causal direction, but the present data do not speak directly to this potential confound. Earlier work did to some extent: when the presentation times of items are controlled by the experimenter, the subjects will prefer the item with longer exposure duration (Armel *et al.*, 2008). However, it is not clear whether the magnitude of the effect is comparable between self-directed and passive fixation. This issue needs to be addressed experimentally.

- We assumed the prior to be Gaussian. Instead, one could allow for a richer parametrization of the prior or use an empirically grounded distribution as the prior. In addition, the agent might update the prior over the course of the experiment.

- Although Step 1 (computing the posterior) and Step 2 (computing the utility difference) of the PUC model are normative, Step 3 (the collapsing bound) is not. One could make this step normative for example by postulating that the agent maximizes expected reward rate (Drugowitsch *et al.*, 2012), but this would make the model quite complicated without clear prospects for additional insight, in part because expected reward rate is only one way to take into account the cost of time.

- One could explore alternative forms of the uncertainty aversion term in Eq. (4.4). We subtracted the standard deviation, but this is somewhat arbitrary. Instead, we could have

subtracted a power of the standard deviation (e.g. variance), or added the inverse standard deviation.

- It is not clear how to generalize our model to the loss domain. For aversive items, the fixation bias seems to be in the opposite direction than for attractive items (Armel *et al.*, 2008); in other words, looking longer at an aversive item makes the item *less* likely to be chosen. An account for this effect could start from the finding that people tend to be risk-seeking in the loss domain (Hershey & Schoemaker, 1980). Replacing risk attitude by uncertainty attitude, it is tempting to simply use $A < 0$ for aversive items in Eq. (4.4) of the model. However, this would not produce a good process model, since it would be ill-specified how the agent sets the sign of A. Instead, we see greater promise in taking a step back and designing an alternative utility function, to replace Eq. (4.4), that is the probability that the item under consideration has a value higher than a criterion v_{crit}:

$$U = \mathbf{Pr}\left(v > v_{\text{crit}} | \mathbf{x}\right). \tag{4.14}$$

  If the posteriors are Gaussian, as in our model, this becomes

$$U = \Phi\left(\mu_{\text{posterior}}; v_{\text{crit}}, \sigma^2_{\text{posterior}}\right), \tag{4.15}$$

  where $\Phi(\cdot;\cdot,\cdot)$ is the cumulative normal distribution with mean and variance parameters. The decision variable would still be the difference of the utilities of the left and right item, as in Eq. (4.9). As an example, we now consider the case of $v_{\text{crit}} = 0$ and $\sigma_p \to \infty$, use the properties of the cumulative normal distribution, and substitute Eqs. (4.7) and (4.8). This yields

$$U = \Phi\left(\frac{\mu_{\text{posterior}}}{\sigma_{\text{posterior}}}; 0, 1\right) = \Phi\left(\frac{\bar{x}\sqrt{T}}{\sigma}; 0, 1\right) \tag{4.16}$$

  For positive v, this U tends to increase with more observations, but for negative v, U tends to decrease. Thus, an aversive item will become less preferred with longer looking time. We conclude that Eq. (4.14) might provide a starting point for future models that generalize better to the loss domain (and that are also less arbitrary in the sense of the previous point).

Finally, we briefly address recent studies that also apply a value inference framework to understand attention-modulated decision-making. These studies interpret the switching of attention as an active sampling process and derive the switching strategy from a optimal policy. The optimal policies are derived in different ways, some with an explicit decreasing threshold like in the PUC model (Song *et al.*, 2019), while others assume that the sampling and switching costs need to be

balanced with accurate posterior estimation (Jang *et al.*, 2021; Callaway *et al.*, 2021). In addition, these models differ in how the behavioral signature of "more fixated item being more preferred" is reproduced. One assumption that they shared is that the prior mean of the item value is either zero or lower than the true value ("prior bias" in Callaway *et al.* 2021). As a result, the less sampled item will have the posterior value closer to prior mean rather than the true value which is higher than the prior. Jang *et al.* (2021) also assumed that attention changes the precision of observations, so that the unattended item will incur samples with bigger variance, thus the expected mean will approach to the true mean even slower; this is similar to the PUC model. Song *et al.* (2019), on the other hand, assumed a distorted value perception for the unattended item by assuming a lower sample mean for that item.

A difference between these approaches and our model is that our reproduction of the qualitative effect does not require an assumption that the prior mean is zero or lower than the true mean; instead, we fix the prior mean to the empirical mean of the ratings on an individual basis. Instead, we introduced the subjective utility with an uncertainty aversion component, the underlying consideration being that consumer decisions may not merely amount to a value comparison similar to perceptual tasks, but also involve choice bias mechanisms. A shortcoming of our approach, however, is that we are not able to predict fixation times. We believe it would be worthwhile to examine whether the PUC model can be equipped with an active sampling mechanism.

At a high level, our work fits in a broader set of recent attempts to appreciate the role of evidence accumulation and inference in value-based decision-making (other examples include Tajima *et al.* 2016; Gabaix & Laibson 2017; Polania *et al.* 2019). We expect the probabilistic inference to become increasingly central in the study of valuation and decision-making.

4.7 Appendix

4.7.1 PUC model variants

We considered three variants of the main PUC model:

- In the "flexible prior variance" model, the prior variance parameter σ_p^2 is a free parameter.
- In the "flexible prior mean & variance" model, both the prior mean μ_p and the prior variance are free parameters.
- In the "uncertainty-neutral" model, the uncertainty aversion parameter A is set to zero, meaning that the utility of an item is equal to the posterior mean of the item's value.

We compared these models to the main PUC model in terms of their log likelihood, AICc, and BIC (Table A). We found that the main PUC model performed best. The uncertainty-neutral model

performed worse than the main model, indicating that the uncertainty aversion term helps to explain people's choices.

To check for potential heterogeneity in the population, we conducted hierarchical Bayesian model selection (Stephan *et al.*, 2009) on the main PUC model and its three variants, using the VBA package (Daunizeau *et al.*, 2014). This analysis indicated that by far the largest proportion of the population follows the main PUC model (97.4% based on AICc and 94.9% based on BIC).

	flexible prior var	flexible prior mean & var	uncertainty-neutral
neg LL	19 (34, 9)	49 (71, 33)	-655 (-455, -923)
AICc	-56 (-75, -25)	-92 (-124, -47)	-1219 (-1756, -835)
BIC	-140 (-159, -108)	-256 (-288, -212)	-1133 (-1666, -741)

Table 4.2: omparison between the main PUC model and its variants, in terms of differences in negative log likelihood, AICc, and BIC. Negative values mean that the main PUC model is better. (Of course, the log likelihood of a more flexible model will always be higher.) All values are summed across subjects, with bootstrapped 95% confidence intervals given in parentheses.

4.7.2 Details of model fitting

4.7.2.1 Parameter ranges

In parameter fitting using Bayesian Adaptive Direct Search, one has to set a range for each parameter that is fitted. A range that is too small may cause the true optimum to fall outside the range. A range that is too large may slow down the optimization or increase the risk of local optima. We set the parameter ranges by trial and error, always ensuring that the fitted parameters did not reach the bounds of the range.

The parameter ranges in the PUC model were as follows:

- The measurement variance σ^2 is always positive. We assigned an upper bound of 900 and a lower bound of e^{-10}.
- For the uncertainty aversion parameter A, we chose a range from -10 to 32.
- The collapsing bound parameters B_0, λ, and k were all restricted to be positive, and we gave them a lower bound of e^{-10}. We choose upper bounds of 100 for B_0 and λ and 20 for k.
- The guessing rate g is a probability and therefore between 0 to 1. In the Weibull function, we used parameter values $a = 21.14$ and $b = 1.38$ in the *wblcdf* function in MATLAB.

- We set non-decision time τ to 0, 100, 200 or 300 ms, given that the temporal resolution of the fixation data that we fitted was 100 ms.

The parameter ranges in the aDDM were as follows:

- All parameters had a lower bound of e^{-10}.
- We gave the scaling constant d an upper bound of 0.1.
- Instead of σ directly, we fitted the scaled quantity $\mu = \frac{d}{\sigma}$, which we assigned an upper bound of 150.
- The attentional bias parameter θ had a upper bound of 1.5.
- The guessing rate and non-decision time parameters had the same ranges as in the PUC model.

For the parameters in the acbDDM that are shared with the aDDM, we used the same parameter ranges as in the aDDM. For the boundary parameters in the acbDDM, we gave k an upper bound of 20 and λ an upper bound of 2000, with a lower bound of e^{-10} for both.

4.7.2.2 Multi-start

When fitting parameters, it is typically a good idea to try multiple random starting points, in order to increase the chance of finding the global optimum. But how to choose the number of starting points? For a somewhat informed choice of this number, we estimated the regret of our fit using a method suggested by Acerbi et al. (see supplement of Acerbi *et al.* 2018):

1. Regret is a function of two positive integers M and N, with $M < N$.
2. Choose N random starting points and run the parameter optimization for each of them. This produces N log likelihood values.
3. Randomly sample, with replacement, M of these N values and calculate their maximum. This simulates the log likelihood that we would have obtained when using M starting points.
4. To reduce noise, repeat the subset sampling and average the results; we used 50 repetitions.
5. The *regret* associated with the pair (M, N) is the maximum of the N original log likelihood values minus the average of the "subset maximum" log likelihood values.

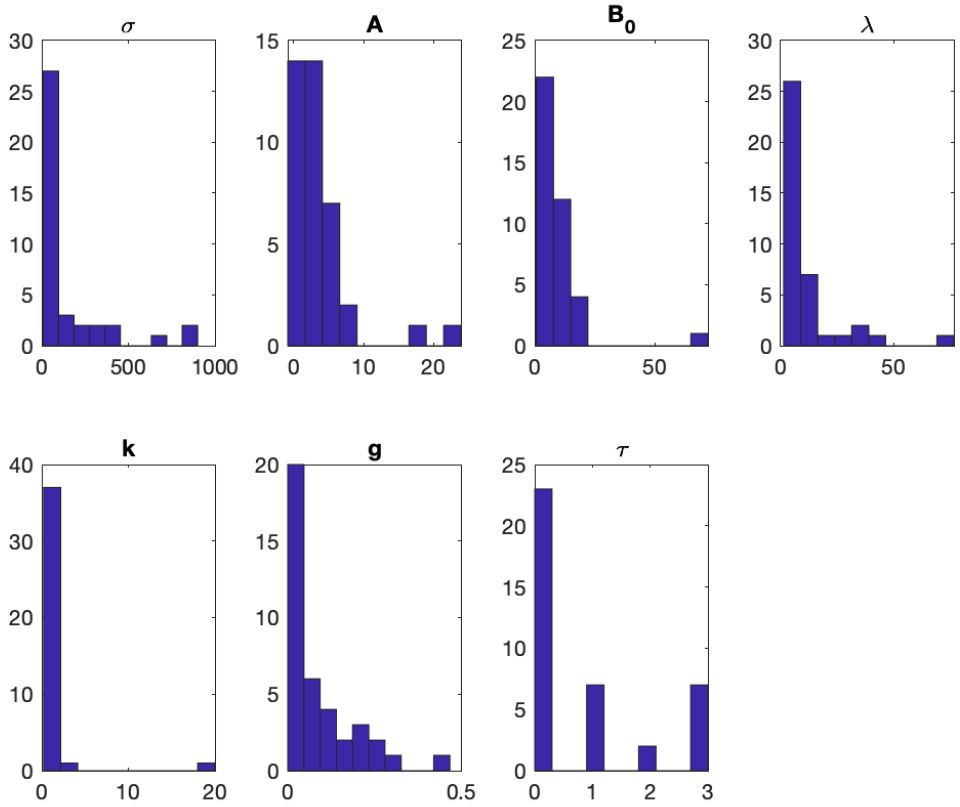


Figure 4.4: Distribution of parameter estimates in the PUC model. σ is the standard deviation of the measurement noise; A is the uncertainty aversion parameter; B_0, λ and k parameterize the collapsing bound function; g is the guessing rate; and τ is the non-decision time.

Finally, we declare a number of starting points N to be *acceptable* if the regret associated with $(N-10, N)$ is no greater than 1. For example, 50 starting points are acceptable if the regret associated with $(40, 50)$ is no greater than 1. This yields a lowest acceptable number of starting points of 23 for the PUC model, 83 for the aDDM, and 91 for the acbDDM. We used these numbers when fitting these models.

4.7.2.3 Parameter estimates

In figure 4.4, 4.5 and 4.6 we show the fitted parameter distribution for the PUC, aDDM and acb-DDM models.

4.7.3 Parameter recovery

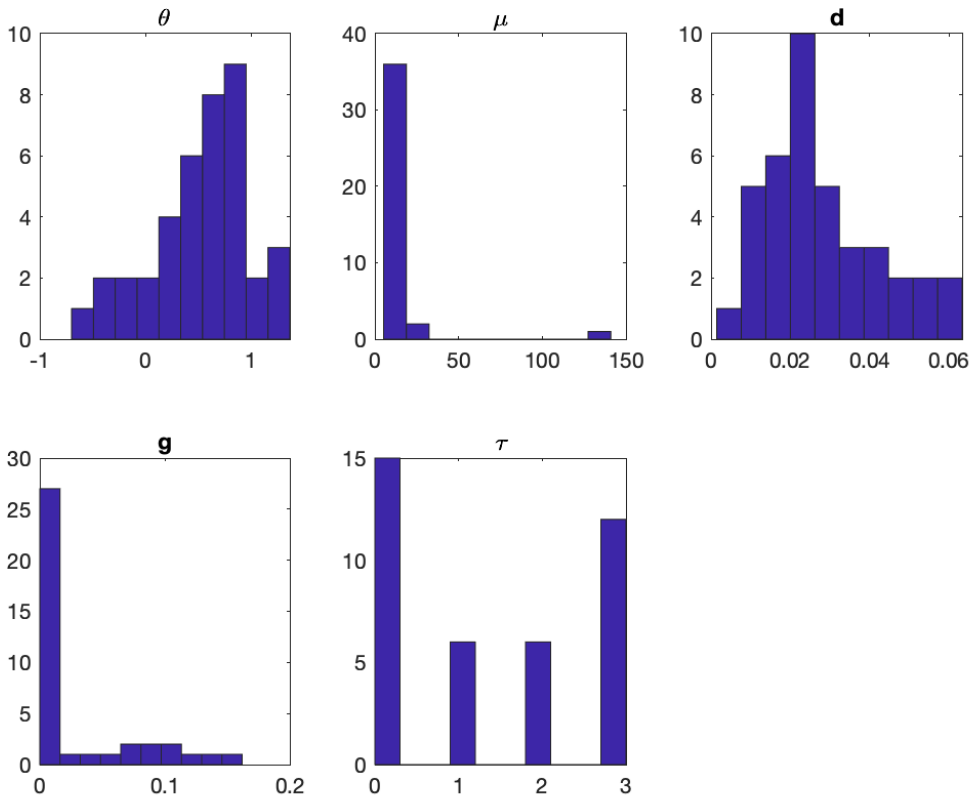


Figure 4.5: Distribution of parameter estimates in the aDDM. θ is the attentional bias factor; μ is the standard deviation of the noise; d is the scaling factor for the decision variable; g is the guessing rate; and τ is the non-decision time.

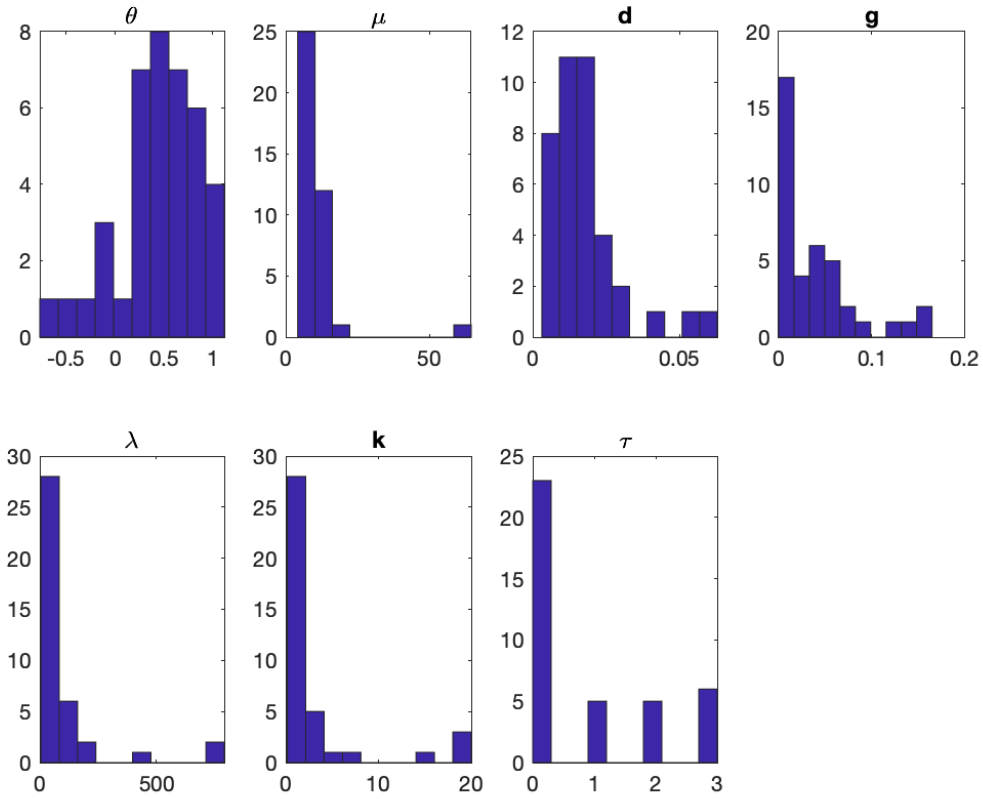
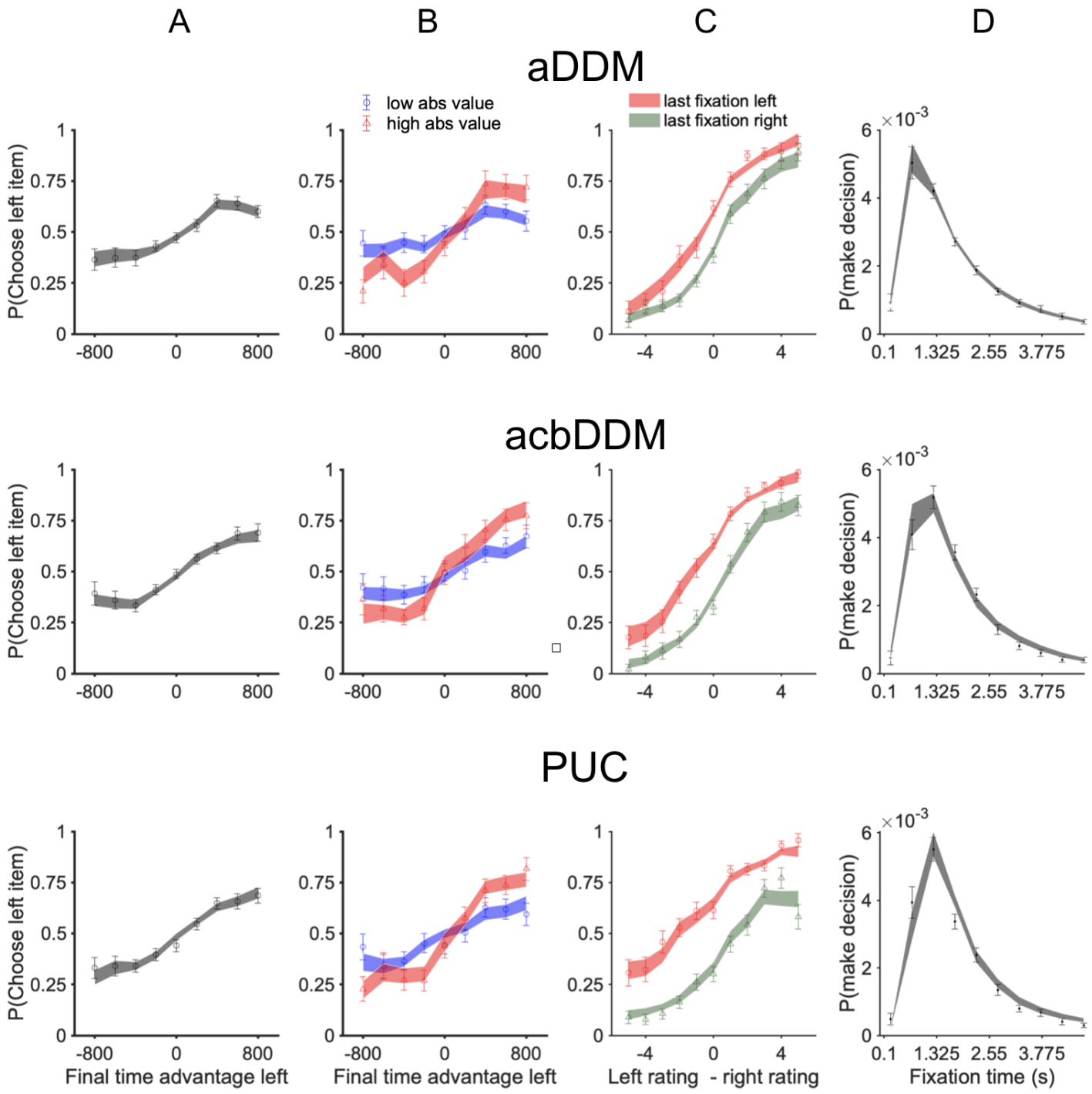


Figure 4.6: Distribution of parameter estimates in the acbDDM. Parameters are as for the aDDM and in addition, λ and k parameterize the collapsing bound.

Within each model, we performed parameter recovery. We generated and then fitted synthetic data sets using the parameter estimates of the real subjects (for a total of 39 data sets). The model fits to the summary statistics obtained using the fitted parameters were near-perfect (Figure **??**.

We then considered the parameter estimates themselves. Parameter recovery was good for the aDDM (Figure 4.7), and less good for the more complex acbDDM and PUC models, with some parameters being recovered well and some less well (Figures 4.8 and 4.9). This is likely due to trade-offs between parameters, where a change in one parameter can be compensated for by changes in one or more other parameters, to produce an approximately equally high log likelihood. Importantly, however, the results of our paper do not rely on parameter estimates but only on maximal log likelihoods (and AICc/BIC), so issues with parameter recovery will not affect our conclusions.



Fits of the aDDM, acbDDM and the PUC models to summary statistics of synthetic data generated from the same respective models.

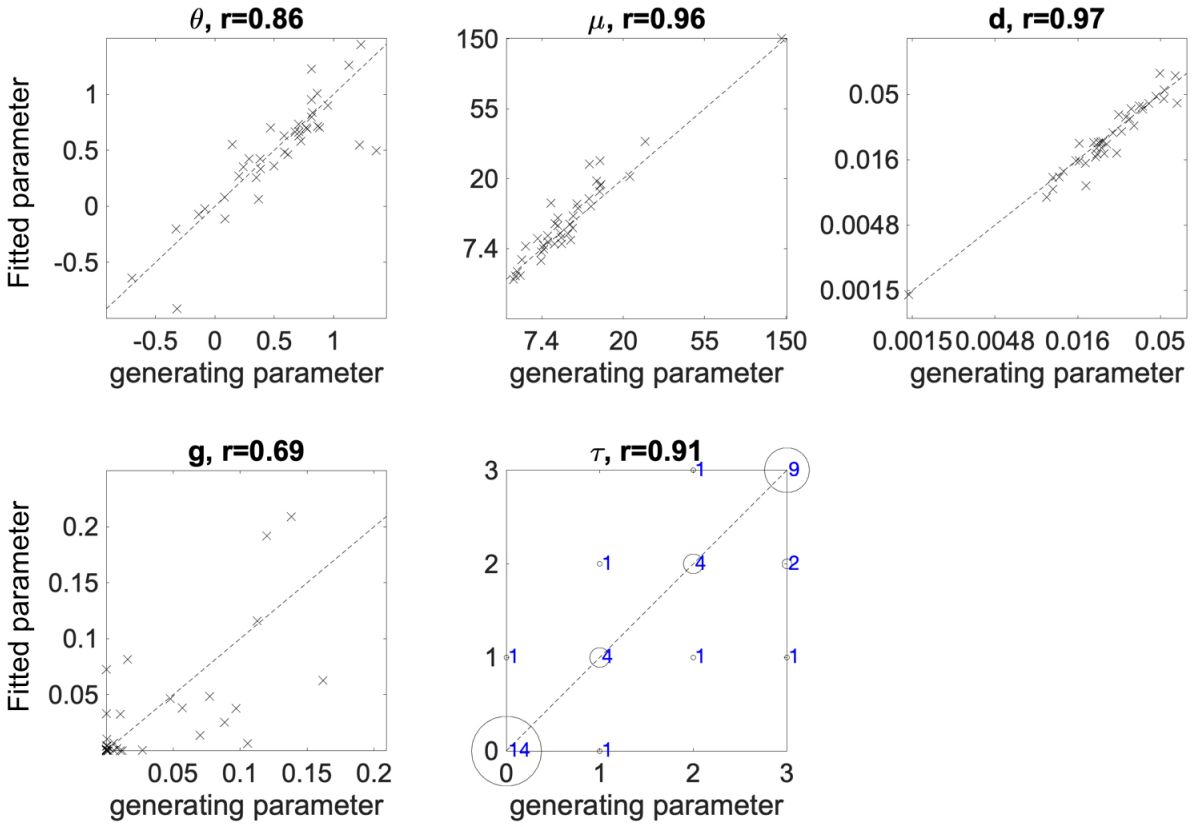


Figure 4.7: Parameter recovery in the aDDM. Shown are the parameter estimates as a function of the generating parameters, with the parameter name and Pearson correlation given in the plot title. Some parameters are plotted in log space because that is how we fitted them; the Pearson correlation is then also calculated for the log parameter. For the non-decision time parameter τ, which has discrete values, circle circumference is proportional to the number of data points as annotated.

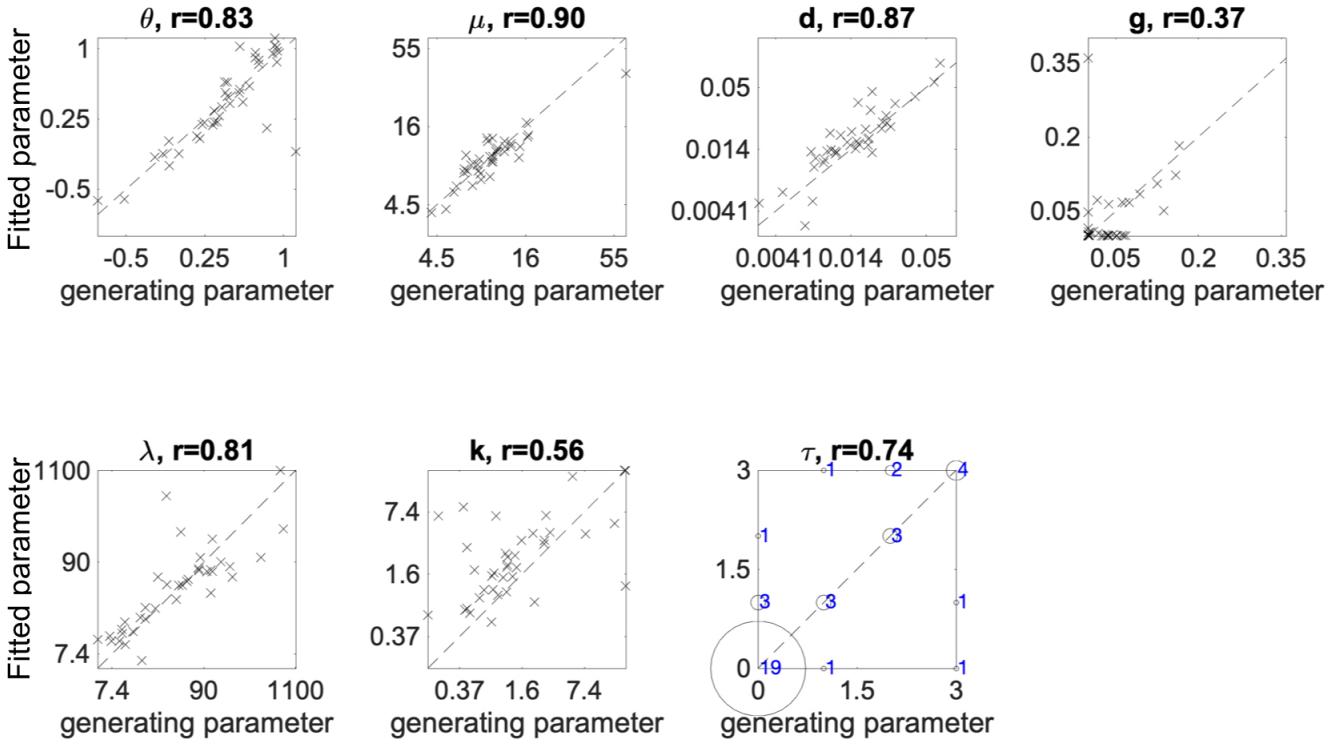


Figure 4.8: Parameter recovery in the acbDDM. For details, see Figure 4.7.

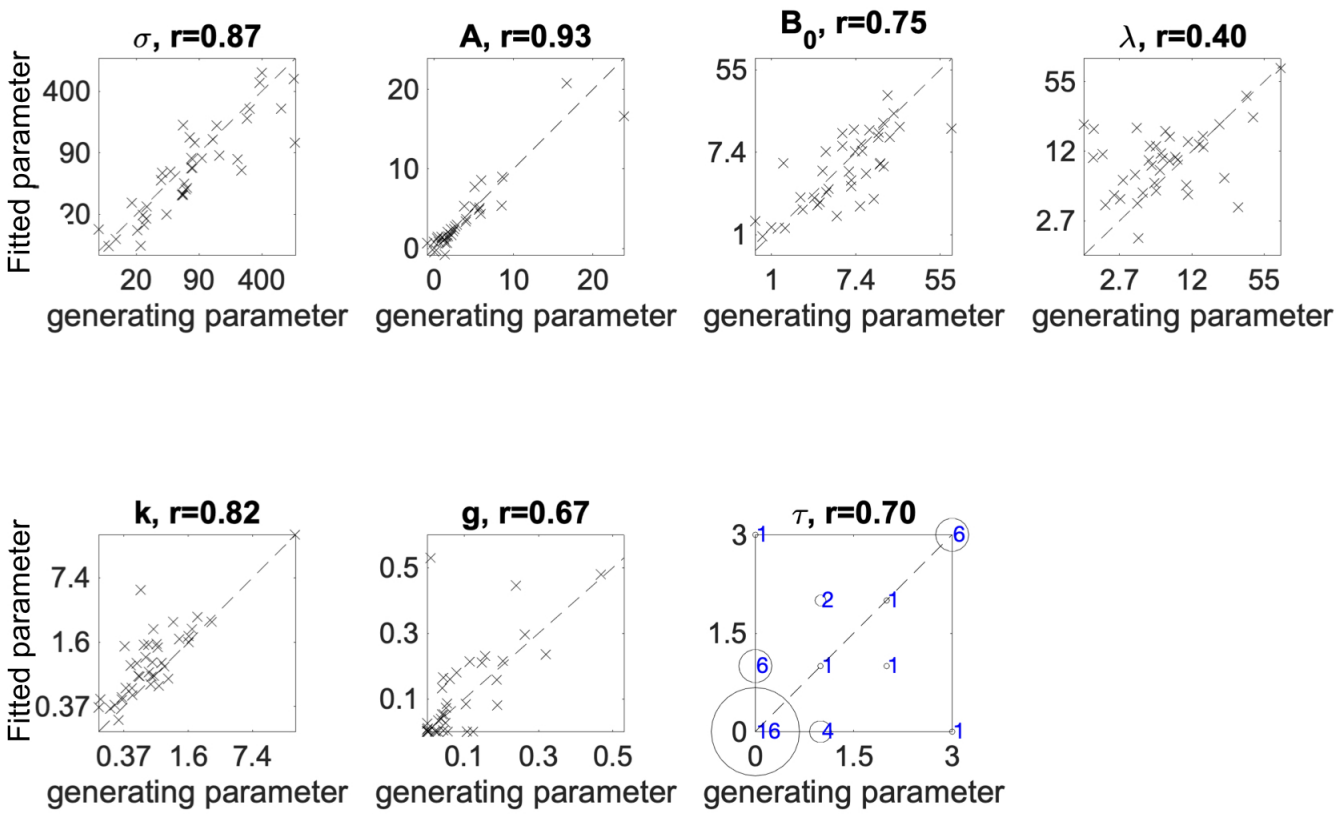


Figure 4.9: Parameter recovery in the PUC model. For details, see Figure 4.7.

BIBLIOGRAPHY

Acerbi, Luigi, & Ma, Wei Ji. 2017. Practical Bayesian optimization for model fitting with Bayesian adaptive direct search. *Pages 1834–1844 of: Proceedings of the 31st International Conference on Neural Information Processing Systems*.

Acerbi, Luigi, Dokka, Kalpana, Angelaki, Dora E, & Ma, Wei Ji. 2018. Bayesian comparison of explicit and implicit causal inference strategies in multisensory heading perception. *PLoS computational biology*, **14**(7), e1006110.

Akaike, Hirotugu. 1974. A new look at the statistical model identification. *IEEE transactions on automatic control*, **19**(6), 716–723.

Alwitt, Linda F. 2002. Suspense and advertising responses. *Journal of Consumer Psychology*, **12**(1), 35–49.

Armel, K. Carrie, Beaumel, Aurelie, & Rangel, Antonio. 2008. Biasing simple choices by manipulating relative visual attention. *Judgment and Decision Making*, **3**(5), 396.

Armeni, Kristijan, Willems, Roel M., & Frank, Stefan L. 2017. Probabilistic language models in cognitive neuroscience: Promises and pitfalls. *Neuroscience & Biobehavioral Reviews*, **83**(Dec.), 579–588.

Baker, Alan. 2016. Simplicity. *In:* Zalta, Edward N. (ed), *The Stanford Encyclopedia of Philosophy*, Winter 2016 edn. Metaphysics Research Lab, Stanford University.

Battaglia, Peter W, Hamrick, Jessica B, & Tenenbaum, Joshua B. 2013. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, **110**(45), 18327–18332.

Bechara, Antoine, & Damasio, Antonio R. 2005. The somatic marker hypothesis: A neural theory of economic decision. *Games and Economic Behavior*, **52**(2), 336–372.

Berlyne, D. E. 1960. *Conflict, arousal, and curiosity.* New York, NY, US: McGraw-Hill Book Company.

Berlyne, D. E. 1966. Curiosity and Exploration. *Science*, **153**(3731), 25–33.

Bezdek, Matthew A, Gerrig, Richard J, Wenzel, William G, Shin, Jaemin, Revill, K Pirog, & Schumacher, Eric H. 2015. Neural evidence that suspense narrows attentional focus. *Neuroscience*, **303**, 338–345.

Bonawitz, Elizabeth Baraff, & Lombrozo, Tania. 2012. Occam's rattle: Children's use of simplicity and probability to constrain inference. *Developmental Psychology*, **48**(4), 1156–1164. Place: US Publisher: American Psychological Association.

Bornstein, Robert F. 1989. Exposure and affect: overview and meta-analysis of research, 1968–1987. *Psychological bulletin*, **106**(2), 265.

Bürkner, Paul-Christian. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of statistical software*, **80**(1), 1–28.

Burnham, Kenneth P, & Anderson, David R. 2004. Multimodel inference: understanding AIC and BIC in model selection. *Sociological methods & research*, **33**(2), 261–304.

Callaway, Frederick, Rangel, Antonio, & Griffiths, Thomas L. 2021. Fixation patterns in simple choice reflect optimal information sampling. *PLoS computational biology*, **17**(3), e1008863.

Carpenter, Bob, Gelman, Andrew, Hoffman, Matthew D, Lee, Daniel, Goodrich, Ben, Betancourt, Michael, Brubaker, Marcus, Guo, Jiqiang, Li, Peter, & Riddell, Allen. 2017. Stan: A probabilistic programming language. *Journal of statistical software*, **76**(1), 1–32.

Cavanagh, James F, Wiecki, Thomas V, Kochar, Angad, & Frank, Michael J. 2014. Eye tracking and pupillometry are indicators of dissociable latent decision processes. *Journal of Experimental Psychology: General*, **143**(4), 1476.

Cheng, Patricia W. 1997. From covariation to causation: A causal power theory. *Psychological review*, **104**(2), 367.

Chowdhary, Kamaljit, & Dupuis, Paul. 2013. Distinguishing and integrating aleatoric and epistemic variation in uncertainty quantification. *ESAIM: Mathematical Modelling and Numerical Analysis*, **47**(3), 635–662.

Churchland, Anne K, Kiani, Roozbeh, & Shadlen, Michael N. 2008. Decision-making with multiple alternatives. *Nature neuroscience*, **11**(6), 693–702.

Comisky, Paul, & Bryant, Jennings. 1982. Factors Involved in Generating Suspense. *Human Communication Research*, **9**(1), 49–58.

Craig, AD. 2003. Interoception: the sense of the physiological condition of the body. *Current opinion in neurobiology*, **13**(4), 500–505.

Cupchtk, Gerald C, Oatleyb, Keith, & Vorderee, Peter. 1998. Emotional effects of reading excerpts from short stories by James Joyce. 15.

Dabney, Will, Kurth-Nelson, Zeb, Uchida, Naoshige, Starkweather, Clara Kwon, Hassabis, Demis, Munos, Rémi, & Botvinick, Matthew. 2020. A distributional code for value in dopamine-based reinforcement learning. *Nature*, 1–5.

Daunizeau, Jean, Adam, Vincent, & Rigoux, Lionel. 2014. VBA: a probabilistic treatment of nonlinear models for neurobiological and behavioural data. *PLoS computational biology*, **10**(1).

de Laplace, P.S. 1820. *Théorie analytique des probabilités*. Courcier. Google-Books-ID: cjAVAAAAQAAJ.

Dearden, Richard, Friedman, Nir, & Russell, Stuart. 1998. Bayesian Q-learning. *Pages 761–768 of: Aaai/iaai*.

Debiec, Jacek, & Olsson, Andreas. 2017. Social Fear Learning: from Animal Models to Human Function. *Trends in Cognitive Sciences*, **21**(7), 546–555.

Drugowitsch, Jan, Moreno-Bote, Rubén, Churchland, Anne K, Shadlen, Michael N, & Pouget, Alexandre. 2012. The cost of accumulating evidence in perceptual decision making. *Journal of Neuroscience*, **32**(11), 3612–3628.

Elpidorou, Andreas. 2018. The bored mind is a guiding mind: toward a regulatory theory of boredom. *Phenom Cogn Sci*, **17**(3), 455–484.

Ely, Jeffrey, Frankel, Alexander, & Kamenica, Emir. 2015. Suspense and surprise. *Journal of Political Economy*, **123**(1), 215–260.

Friston, Karl J., & Stephan, Klaas E. 2007. Free-energy and the brain. *Synthese*, **159**(3), 417–458.

Gabaix, Xavier, & Laibson, David. 2017. *Myopia and Discounting [Working Paper]*. Tech. rept.

Gerrig, Richard J., & Bernardo, Allan B. I. 1994. Readers as problem-solvers in the experience of suspense. *Poetics*, **22**(6), 459–472.

Golkar, Armita, Castro, Vasco, & Olsson, Andreas. 2015. Social learning of fear and safety is determined by the demonstrator's racial group. *Biology Letters*, **11**(1), 20140817.

Gottlieb, Jacqueline, & Oudeyer, Pierre-Yves. 2018. Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, **19**(12), 758–770.

Griffiths, Thomas L., & Tenenbaum, Joshua B. 2009. Theory-based causal induction. *Psychological Review*, **116**(4), 661–716.

Gruber, Matthias J., & Ranganath, Charan. 2019. How Curiosity Enhances Hippocampus-Dependent Memory: The Prediction, Appraisal, Curiosity, and Exploration (PACE) Framework. *Trends in Cognitive Sciences*, **23**(12), 1014–1025.

Gureckis, Todd M, Martin, Jay, McDonnell, John, Rich, Alexander S, Markant, Doug, Coenen, Anna, Halpern, David, Hamrick, Jessica B, & Chan, Patricia. 2016. psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior research methods*, **48**(3), 829–842.

Gwinn, Rachael, & Krajbich, Ian. 2020. Attitudes and attention. *Journal of Experimental Social Psychology*, **86**, 103892.

Harrison, Glenn W, & Elisabet Rutström, E. 2008. Risk aversion in the laboratory. *Pages 41–196 of: Risk aversion in experiments*. Emerald Group Publishing Limited.

Hawkins, G. E., Forstmann, B. U., Wagenmakers, E.-J., Ratcliff, R., & Brown, S. D. 2015. Revisiting the Evidence for Collapsing Boundaries and Urgency Signals in Perceptual Decision-Making. *Journal of Neuroscience*, **35**(6), 2476–2484.

Hershey, John C, & Schoemaker, Paul JH. 1980. Risk taking and problem context in the domain of losses: An expected utility analysis. *Journal of Risk and Insurance*, 111–132.

Horan, Mattias, Daddaoua, Nabil, & Gottlieb, Jacqueline. 2019. Parietal neurons encode information sampling based on decision uncertainty. *Nature Neuroscience*, **22**(8), 1327–1335.

Jang, Anthony Injoon, Sharma, Ravi, & Drugowitsch, Jan. 2021. Optimal policy for attention-modulated decisions explains human fixation behavior. *Elife*, **10**, e63436.

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

Kang, Min Jeong, Hsu, Ming, Krajbich, Ian M., Loewenstein, George, McClure, Samuel M., Wang, Joseph Tao-yi, & Camerer, Colin F. 2009. The Wick in the Candle of Learning: Epistemic Curiosity Activates Reward Circuitry and Enhances Memory. *Psychological Science*, **20**(8), 963–973.

Kaufmann, Nicolas, Schulze, Thimo, & Veit, Daniel. 2011. More than fun and money. Worker Motivation in Crowdsourcing-A Study on Mechanical Turk. *Pages 1–11 of: Amcis*, vol. 11. Detroit, Michigan, USA.

Klimmt, Christoph, Rizzo, Albert, Vorderer, Peter, Koch, Jan, & Fischer, Till. 2009. Experimental Evidence for Suspense as Determinant of Video Game Enjoyment. *CyberPsychology & Behavior*, **12**(1), 29–31.

Knill, David C, & Richards, Whitman. 1996. *Perception as Bayesian inference*. Cambridge University Press.

Knobloch-Westerwick, Silvia, David, Prabu, Eastin, Matthew S., Tamborini, Ron, & Greenwood, Dara. 2009. Sports Spectators' Suspense: Affect and Uncertainty in Sports Entertainment. *Journal of Communication*, **59**(4), 750–767.

Krajbich, Ian. 2019. Accounting for attention in sequential sampling models of decision making. *Current opinion in psychology*, **29**, 6–11.

Krajbich, Ian, & Rangel, Antonio. 2011. Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, **108**(33), 13852–13857.

Krajbich, Ian, & Smith, Stephanie M. 2015. Modeling Eye Movements and Response Times in Consumer Choice. *Journal of Agricultural & Food Industrial Organization*, **13**(1).

Krajbich, Ian, Armel, Carrie, & Rangel, Antonio. 2010. Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, **13**(10), 1292–1298.

Krajbich, Ian, Lu, Dingchao, Camerer, Colin, & Rangel, Antonio. 2012. The Attentional Drift-Diffusion Model Extends to Simple Purchasing Decisions. *Frontiers in Psychology*, **3**(June).

LeDoux, Joseph E. 2014. Coming to terms with fear. *PNAS*, **111**(8), 2871–2878. Publisher: National Academy of Sciences Section: Biological Sciences.

LeDoux, Joseph E., & Brown, Richard. 2017. A higher-order theory of emotional consciousness. *PNAS*, **114**(10), E2016–E2025. Publisher: National Academy of Sciences Section: PNAS Plus.

Lee, Angela Y. 2001. The mere exposure effect: An uncertainty reduction explanation revisited. *Personality and Social Psychology Bulletin*, **27**(10), 1255–1266.

Lehne, Moritz, & Koelsch, Stefan. 2015. Toward a general psychological model of tension and suspense. *Frontiers in Psychology*, **6**.

Levy, Haim, & Markowitz, Harry M. 1979. Approximating expected utility by a function of mean and variance. *The American Economic Review*, **69**(3), 308–317.

Lim, Jonathan B, & Oppenheimer, Daniel M. 2020. Explanatory preferences for complexity matching. *PloS one*, **15**(4), e0230929.

Loewenstein, George. 1994. The psychology of curiosity: A review and reinterpretation. *Psychological bulletin*, **116**(1), 75.

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

Lu, Hongjing, Yuille, Alan L., Liljeholm, Mimi, Cheng, Patricia W., & Holyoak, Keith J. 2008. Bayesian generic priors for causal learning. *Psychological Review*, **115**(4), 955–984.

Ma, Wei Ji. 2010. Signal detection theory, uncertainty, and Poisson-like population codes. *Vision Research*, **50**(22), 2308 – 2319. Mathematical Models of Visual Coding.

Ma, Wei Ji, & Jazayeri, Mehrdad. 2014. Neural Coding of Uncertainty and Probability. *Annual Review of Neuroscience*, **37**(1), 205–220.

Madumal, Prashan, Miller, Tim, Sonenberg, Liz, & Vetere, Frank. 2020. Explainable reinforcement learning through a causal lens. *Pages 2493–2500 of: Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34.

Maren, Stephen. 2001. Neurobiology of Pavlovian Fear Conditioning. *Annual Review of Neuroscience*, **24**(1), 897–931.

Markowitz, Harry. 1952. Portfolio selection. *The journal of finance*, **7**(1), 77–91.

Milosavljevic, Milica, Malmaud, Jonathan, Huth, Alexander, Koch, Christof, & Rangel, Antonio. 2010. The Drift Diffusion Model can account for the accuracy and reaction time of value-based choices under high and low time pressure. *Judgment and Decision Making*, **5**(6), 437–449.

Myrvold, Wayne C. 2003. A Bayesian account of the virtue of unification. *Philosophy of Science*, **70**(2), 399–423.

Nelson, Jonathan D. 2005. Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological review*, **112**(4).

Nelson, Jonathan D., McKenzie, Craig R.M., Cottrell, Garrison W., & Sejnowski, Terrence J. 2010. Experience Matters: Information Acquisition Optimizes Probability Gain. **21**(7), 960–969. Publisher: SAGE Publications Inc.

Nomikos, Markellos S., Opton Jr., Edward, & Averill, James R. 1968. Surprise versus suspense in the production of stress reaction. *Journal of Personality and Social Psychology*, **8**(2, Pt.1), 204–208.

Olsson, Andreas, & Phelps, Elizabeth A. 2007. Social learning of fear. *Nature Neuroscience*, **10**(9), 1095–1102.

Pacer, M, & Lombrozo, Tania. 2017. Ockham's razor cuts to the root: Simplicity in causal explanation. *Journal of Experimental Psychology: General*, **146**(12), 1761–1780.

Pacer, Michael, Williams, Joseph, Chen, Xi, Lombrozo, Tania, & Griffiths, Thomas. 2013. Evaluating computational models of explanation using human judgments. *arXiv:1309.6855 [cs]*, Sept. arXiv: 1309.6855.

Pärnamets, Philip, Johansson, Petter, Hall, Lars, Balkenius, Christian, Spivey, Michael J, & Richardson, Daniel C. 2015. Biasing moral decisions by exploiting the dynamics of eye gaze. *Proceedings of the National Academy of Sciences*, **112**(13), 4170–4175.

Pavlov, P Ivan. 2010. Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex. *Annals of neurosciences*, **17**(3), 136.

Pearl, Judea. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

Peterson, Erik M., & Raney, Arthur A. 2008. Reconceptualizing and Reexamining Suspense as a Predictor of Mediated Sports Enjoyment. *Journal of Broadcasting & Electronic Media*, **52**(4), 544–562.

Polania, Rafael, Woodford, Michael, & Ruff, Christian C. 2019. Efficient coding of subjective value. *Nature neuroscience*, **22**(1), 134.

Pouget, Alexandre, Beck, Jeffrey M, Ma, Wei Ji, & Latham, Peter E. 2013. Probabilistic brains: knowns and unknowns. *Nature Neuroscience*, **16**(9), 1170–1178.

Ratcliff, Roger, & McKoon, Gail. 2008. The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation*, **20**(4), 873–922.

Scherer, Klaus R. 2001. Appraisal considered as a process of multilevel sequential checking. *Appraisal processes in emotion: Theory, methods, research*, **92**(120), 57.

Schwarz, Gideon, *et al.* 1978. Estimating the dimension of a model. *The annals of statistics*, **6**(2), 461–464.

Seabold, Skipper, & Perktold, Josef. 2010. statsmodels: Econometric and statistical modeling with python. *In: 9th Python in Science Conference*.

Shadlen, Michael N, & Shohamy, Daphna. 2016. Decision making and sequential sampling from memory. *Neuron*, **90**(5), 927–939.

Shimojo, Shinsuke, Simion, Claudiu, Shimojo, Eiko, & Scheier, Christian. 2003. Gaze bias both reflects and influences preference. *Nature Neuroscience*, **6**(12), 1317–1322.

Smith, Stephanie M, & Krajbich, Ian. 2019. Gaze amplifies value in decision making. *Psychological science*, **30**(1), 116–128.

Song, Mingyu, Wang, Xingyu, Zhang, Hang, & Li, Jian. 2019. Proactive information sampling in value-based decision-making: deciding when and where to saccade. *Frontiers in human neuroscience*, **13**, 35.

Stephan, Klaas Enno, Penny, Will D., Daunizeau, Jean, Moran, Rosalyn J., & Friston, Karl J. 2009. Bayesian model selection for group studies. *NeuroImage*, **46**(4), 1004–1017.

Su-lin, Gan, Tuggle, Charles A, Mitrook, Michael A, Coussement, Sylvere H, & Zillmann, Dolf. 1997. The thrill of a close game: Who enjoys it and who doesn't? *Journal of Sport and Social Issues*, **21**(1), 53–64.

Tajima, Satohiro, Drugowitsch, Jan, & Pouget, Alexandre. 2016. Optimal policy for value-based decision-making. *Nature communications*, **7**, 12400.

Tavares, Gabriela, Perona, Pietro, & Rangel, Antonio. 2017. The attentional drift diffusion model of simple perceptual decision-making. *Frontiers in neuroscience*, **11**, 468.

Thagard, Paul. 1989. Explanatory coherence. *Behavioral and brain sciences*, **12**(3), 435–467.

Tversky, Amos, & Kahneman, Daniel. 1983. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review*, **90**(4), 293.

Vallverdú, Jordi, & Giannoccaro, Ivan (eds). 2015. *Handbook of Research on Synthesizing Human Emotion in Intelligent Systems and Robotics:*. Advances in Computational Intelligence and Robotics. IGI Global.

van Lieshout, Lieke LF, Vandenbroucke, Annelinde RE, Müller, Nils CJ, Cools, Roshan, & de Lange, Floris P. 2018. Induction and relief of curiosity elicit parietal and frontal activity. *Journal of Neuroscience*, **38**(10), 2579–2588.

van Opheusden, Bas, Acerbi, Luigi, & Ma, Wei Ji. 2020. Unbiased and efficient log-likelihood estimation with inverse binomial sampling. *PLoS computational biology*, **16**(12), e1008483.

von Haugwitz, Rickard, & Dodig-Crnkovic, Gordana. 2015. Probabilistic Computation and Emotion as Self-regulation. *Pages 1–4 of: Proceedings of the 2015 European Conference on Software Architecture Workshops*. Dubrovnik Cavtat Croatia: ACM.

Walker, Edgar Y., Cotton, R. James, Ma, Wei Ji, & Tolias, Andreas S. 2020. A neural basis of probabilistic computation in visual cortex. *Nat Neurosci*, **23**(1), 122–129. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Decision;Neural decoding;Neural encoding;Sensory processing;Striate cortex Subject_term_id: decision;neural-decoding;neural-encoding;sensory-processing;striate-cortex.

Webb, Ryan. 2019. The (neural) dynamics of stochastic choice. *Management Science*, **65**(1), 230–255.

Westbrook, Andrew, van den Bosch, R, Määttä, JI, Hofmans, L, Papadopetraki, D, Cools, Roshan, & Frank, MJ. 2020. Dopamine promotes cognitive effort by biasing the benefits versus costs of cognitive work. *Science*, **367**(6484), 1362–1366.

White, J. Kael, & Monosov, Ilya E. 2016. Neurons in the primate dorsal striatum signal the uncertainty of object–reward associations. *Nature Communications*, **7**(1), 1–8.

White, J. Kael, Bromberg-Martin, Ethan S., Heilbronner, Sarah R., Zhang, Kaining, Pai, Julia, Haber, Suzanne N., & Monosov, Ilya E. 2019. A neural network for information seeking. *bioRxiv*, Aug., 720433.

Williams, David. 1991. *Probability with martingales*. Cambridge university press.

Wilmot, David, & Keller, Frank. 2020. Suspense in Short Stories is Predicted By Uncertainty Reduction over Neural Story Representation. *Pages 1763–1788 of: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Yanal, Robert J. 1996. The paradox of suspense. *The British Journal of Aesthetics*, **36**(2), 146–159.

Yuan, Changhe, Lim, Heejin, & Lu, Tsai-Ching. 2011. Most relevant explanation in Bayesian networks. *Journal of Artificial Intelligence Research*, **42**, 309–352.

Zajonc, Robert B. 2001. Mere exposure: A gateway to the subliminal. *Current directions in psychological science*, **10**(6), 224–228.

Zemla, Jeffrey C., Sloman, Steven, Bechlivanidis, Christos, & Lagnado, David A. 2017. Evaluating everyday explanations. *Psychon Bull Rev*, **24**(5), 1488–1500.

Zemla, Jeffrey C, Sloman, Steven A., Bechlivanidis, Christos, & Lagnado, David. 2020 (Jan.). *Not so simple! Mechanisms increase preference for complex explanations*. preprint. PsyArXiv.

Zillmann, Dolf. 1991. The logic of suspense and mystery. *Responding to the screen: Reception and reaction processes*, **7**, 281–303.