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.