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

© Zhiwei Li 2021

TABLE OF CONTENTS

LIST OF FIGURES iv

LIST OF TABLES viii

ABSTRACT x

CHAPTER

1 Introduction 1

1.1 modeling suspense as expected future learning 1
1.2 modeling satisfying explanation as a combination of distinctive causes 2
1.3 modeling value-based choice as maximizing a posterior based utility function . . 3

2 Expectations about future learning influence moment-to-moment feelings of suspense 5

2.1 abstract 5
2.2 Introduction 5
2.2.1 Previous research 6
2.2.2 Theory: Suspense as the expected future belief update 9
2.3 An experimental test of the theory 12
2.3.1 Overview of the experiments 13
2.4 Experiment 1: Predicting the dynamics of suspense 13
2.4.1 Experimental Method 13
2.4.2 Modeling Approach 17
2.4.3 Alternative models 17
2.4.4 Results 19
2.4.5 Discussion 22
2.5 Experiment 2: Rules with different temporal structures 23
2.5.1 Method 23
2.5.2 Results 24
2.5.3 Discussion 26
2.6 Experiment 3: Manipulating willingness to play more games 26
2.6.1 Method 26
2.6.2 Results 27
2.6.3 Discussion 28
2.7 General Discussion 28
2.7.1 Evaluation of models across all experiments 29

2.7.2 Limitations and future directions 30
2.8 Appendix 32
2.8.1 Stimuli design 32
2.8.2 Data from three experiments 32
2.8.3 Results from Experiment 3 32
2.8.4 Computing model likelihood 32
2.8.5 Derivation of belief update 34
2.8.6 Alternative models 34

3 Quantifying explanation preference in a probabilistic context 46

3.1 Abstract 46
3.2 Introduction 46
3.3 Research goal 49
3.4 Overview of experimental procedures 50
3.5 Experiment 1 52
3.5.1 Method 53
3.5.2 Results 55
3.5.3 Discussion 61
3.6 Experiment 2 62
3.6.1 Method 63
3.6.2 Results 63
3.6.3 Discussion 72
3.7 General discussion 72
3.8 Appendix 74
3.8.1 Details of the experimental material 74
3.8.2 Additional results 78

4 Choice preference with posterior-based account 82

4.1 Abstract 82
4.2 Introduction 82
4.3 Models 84
4.3.1 Posterior-Utility-Choice (PUC) model 84
4.3.2 Attentional drift-diffusion model 91
4.4 Materials and Methods 93
4.5 Results 95
4.6 Discussion 98
4.7 Appendix 101
4.7.1 PUC model variants 101
4.7.2 Details of model fitting 102
4.7.3 Parameter recovery 104

BIBLIOGRAPHY 112

LIST OF FIGURES

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. 11

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. 15

2.3 Experiment 1, Distribution of reported suspense in low and high predicted-suspense conditions. 21

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 37

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 38

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 39

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) 39

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"). 40

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). 41

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. 42

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. 43

2.12 Illustration of the last three time steps of state transitions. 44

2.13 Comparing different metrics. 44

2.14 Comparing different metrics given the special case of p(now)=0.5 45

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. 50

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. 52

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. 56

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. 56

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. 57

3.6 Choice ratio of the three explanations for trials with equal prior and causal strengths, separated by the two clusters of participants. 60
3.7 Rating of conjunctive priors, separated by the two clusters of participants. 60
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. 62
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. 65
3.10 Empirical data of trials with unequal causal strength and equal prior. A: Averaged ratio of choosing 1-, 2- or 3-cause; error bars indicate the 95% confidence interval calculated from bootstrapping. In the legend, "p" denotes prior and "cs" denotes causal strength of each trial. B: average ratio of choosing simple cause A, separated by the level of prior and causal strength level. C: average ratio of choosing complex cause A+B+C, separated by the level of prior and causal strength level. 67
3.11 Empirical choice probability for trials of equal causal strength but equal or unequal prior. We categorized trials by their prior condition, by whether the trial has equal prior, or the first cause having the highest prior ("high 1 pr" in the figure), or the second having the highest prior ("high 2 pr"). For each trial type, the ratio of choosing one-, two- or three-cause explanations is shown in different colored lines. 68
3.12 Empirical choice probability for trials of unequal causal strength but equal or unequal prior. The first cause always has the highest causal strength. We categorized trials by their prior condition, by whether the trial has equal prior, or the second having the highest prior ("high 2 pr"). For each trial type, the ratio of choosing one- or two- or three-cause explanations is shown in different colored lines. 70
3.13 The likelihood difference between heuristic model and Bayesian model for each participant. 71
3.14 An example trial for the prior judgement section. 76
3.15 An example trial for the causal strength judgement section. 77
3.16 Choice ratio of the three explanations, separated by the two clusters of participants. . . 79
3.17 Rating of conjunctive priors, separated by the two clusters of participants, for Experiment 2. 80
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. 81

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. 85

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). 88

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*). 96

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. 104

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. 105

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. 106

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. 109

4.8 Parameter recovery in the acbDDM. For details, see Figure 4.7. 110

4.9 Parameter recovery in the PUC model. For details, see Figure 4.7. 111

LIST OF TABLES

TABLE

2.1 Alternative Models 20

2.2 Experiment 1: Model Comparison. 21

2.3 Experiment 2: Model Comparison 25

2.4 Model comparison for all Experiments. 29

2.5 Experiment 3: Model Comparison 33

3.1 Comparing the predicted best explanation with empirical average best explanation, where 2 means 2-cause explanation, 3-means 3-cause explanation, and so on. The prediction is generated from the heuristic model that chooses the explanation including all the maximum causal strength factors but nothing extra. The empirically best explanation and the proportion of participants choosing this answer are listed in the last column. 58

3.2 Comparing the predicted best explanation from heuristic model and Bayesian model with empirical average best explanation in the free response data. 59

3.3 Regression results for trials with equal prior and equal causal strength. Dependent variable (DV) are the probability of choosing the explanation “A” or “A+B+C”, denoted as P(A) and P(A+B+C), respectively. Intercept is the average baseline of choosing each option, other rows are the fitted slope regarding the specific regressor. “prior” denotes the prior level of high or low, “cs” denotes causal strength level of high or low, “interaction” denotes whether prior and cs are in the same direction or the opposite. The columns represents estimated average, estimated standard error, lower and higher edge of the 95% confidence interval. If the interval includes 0 that means the regressor is not significant. 64

3.4 Regression for the equal prior and unequal causal strength trials. Note “cs” here denotes causal strength difference. 66

3.5 Comparing the predicted best explanation with empirical average best explanation. The prediction is from the heuristic that chooses the explanation including all the factors of maximum prior and maximum causal strength but not more than that. The empirically best explanation and the proportion of participants choosing this answer are listed in the last column. 69

3.6 Comparing the predicted best explanation with empirical average best explanation for the free response data. “cs heuristic” is the one we proposed in Experiment 1 which only takes into account of causal strength; while the “full heuristic” is the one we presented above. The empirical best comes with the percentage of participants choosing this answer. For the free response, there are in total 7 possible responses therefore the random baseline is 14.8%. 69

3.7 Comparing the predicted best explanation with empirical average best explanation for data in Experiment 1. The "cs heuristic" is the one we proposed in Experiment 1 while the "full heuristic" is the one we presented in the current section. 71
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. 78
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. 79

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. 98
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. 102

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 developed 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.