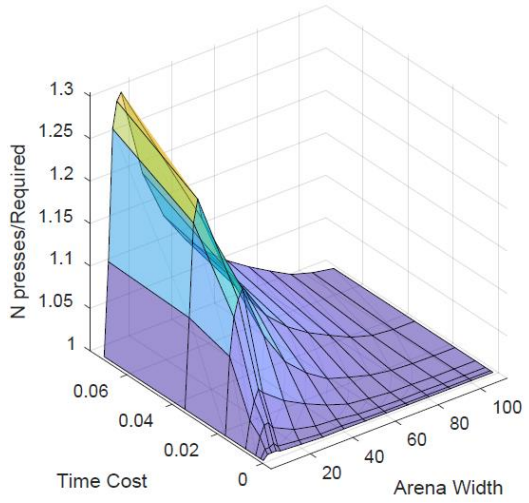
OPTIMAL DECISION MAKING FOR TEMPORALLY EXTENDED ACTIONS

Roberto Medina



Dissertation presented to obtain the Ph.D degree in Neuroscience

Instituto de Tecnologia Química e Biológica António Xavier | Universidade Nova de Lisboa

Oeiras,
November,
2019

INSTITUTO
DE TECNOLOGIA
QUÍMICA E BIOLÓGICA
/UNL

Knowledge Creation

OPTIMAL DECISION MAKING FOR TEMPORALLY EXENDED ACTIONS

Roberto Medina

Dissertation presented to obtain the Ph.D degree in Neuroscience

Instituto de Tecnologia Química e Biológica António Xavier | Universidade Nova de Lisboa

Supervisor: Alfonso Renart PhD

Research work coordinated by:

Oeiras,
November, 2019

INSTITUTO
DE TECNOLOGIA
QUÍMICA E BIOLÓGICA
/UNL

Knowledge Creation

Aknowledgements:

In first place I would like to acknowledge and thank Dr. Alfonso Renart for providing an environment for scientific questioning that nourished my mind and resulted in an investigation beyond anything I had previously envisioned. I would also like to thank my thesis committee, Dr. Leopoldo Petreanu and Dr. Joe Paton, the members of the jury, my collaborators who exponentially potentiated this work, Dr. Gautam Agarwal, Dr. Jose L. Pardo-Vazquez, Juan Castineiras, Miguel Bengala, and Dr. Zach Mainen, Dr. Fatima Husain for believing in me and setting me on this path, my dear friends John Connor and Rami Abuhabsah for graciously letting me crash in their living room at the humble start of my scientific journey, Dr. Luis Moreira for listening and giving me feedback on this work to exhaustion, Dr. Gonçalo Lopes for Bonsai and teaching me how to measure behavior, my lab, my INDP 2011 classmates, and a special thank you to my partner and office neighbor Dr. Simone Lackner for innumerable scientific discussions, her feedback and encouragement at any and all times, and ultimately helping me mature as a scientist.

Obrigado.

For Nidia, Kimberly, Tony, Gloria and Roberto.

Abstract:

The application of the scientific method to the study of behavior has gone beyond the realm of philosophy and psychology, requiring modern neuroscientists to understand neurobiology, statistics, and even computer science. Neurobiology explains that neural systems in all animals, including humans, were evolved to meet specific environmental demands in order for the organism to survive, mate, and produce viable offspring. However, the complexity of studying behavior in natural environments has forced neuroscientists to construct reduced artificial environments where human and non-human animals alike can be trained to respond to calibrated stimuli. For example, a conventional approach to investigate thresholds for discrimination of 2 distinct stimuli has been to ask subjects to press one of two buttons according to the identity of the presented stimulus.

In order to study the process of committing to a choice from online behavior, in this thesis we expand the standard 2-choice psychophysical task, replacing the traditional two response buttons by an arena that must be traversed in order to reach one of two distally separated targets. This modification and the use of time-varying stimuli to signal the correct target allowed us to observe that difficult choices can elicit complex choice trajectories that meander between the two targets in a stimulus correlated manner. We show that this behavior is also performed by a reinforcement learning agent that moves in order to maximize long term rewards while considering movement and time constraints. The agent estimates what she believes to be the correct target by accumulating stimulus evidence, and subtracts from the associated reward for answering correctly a cost for moving and a cost for waiting. The resulting cost function eventually commits the agent to a target, with the peculiar behavior that, when time allows, the agent changes her direction as her belief crosses belief thresholds. We also that show that these bounds on the agent's beliefs explicitly depend on the movement and time costs. We interpret the choice trajectories displayed by the human and the model to be sequences of hesitant choices (crossing soft or reversible decision bounds) in the process of committing to a final choice (crossing a hard or irreversible decision bound).

Resumo:

Para estudar o processo de compromisso com uma escolha, nesta tese, expandimos a tarefa psicofísica convencional de 2 opções, substituindo os dois botões de resposta tradicionais por uma arena que deve ser percorrida para alcançar um dos dois objetivos distalmente separados. Essa modificação e o uso de estímulos que variam no tempo para sinalizar o objetivo correto nos permitiram observar que escolhas difíceis podem desencadear trajetórias complexas de escolha que serpenteiam entre os dois objetivos de maneira correlacionada ao estímulo. Mostramos que esse comportamento também é realizado por um agente simulado que aprende por reforço, que se move para maximizar recompensas de longo prazo, considerando as restrições de movimento e tempo. A agente calcula o que ela acredita ser a meta correta, acumulando evidências de estímulo e subtrai da recompensa associada por responder corretamente um custo de deslocamento no espaço e um custo de espera. A função de custo resultante acaba comprometendo o agente com um objetivo, com o comportamento peculiar de que, quando o tempo permitir, o agente muda de direção à medida que sua crença cruza os limites da crença. Também mostramos que esses limites ás crenças do agente dependem explicitamente dos custos de movimento e tempo. Interpretamos as trajetórias de escolha exibidas pelo humano e o modelo como sequências de escolhas hesitantes (cruzando limites de decisão reversíveis) no processo de se comprometer com uma escolha final (cruzando um limite de decisão irreversível).

Author contributions:

I also start this thesis by acknowledging the collaborators that took part in different phases of this project, and some of the people that took the time to discuss my work in depth. All of this work was done under the guidance and supervision of Alfonso Renart.

- Chapter 2: Experiment 1 data was collected and analyzed by me with the guidance of Jose L. Pardo-Vazquez and Alfonso Renart. For experiment 2, we thank Gautam Agarwal for his contribution in providing an innovative approach to analyzing our data and for discussions, and Jose L. Pardo-Vazquez and Zach Mainen for discussions.

- Chapter 3: Chapter 3 was analyzed by me with the guidance of Alfonso Renart. We would like to thank Gautam Agarwal for providing the first normative model of our task, to Juan Castineiras for his derivation of an analytical equation for the Bayesian belief propagation driving reward expectation in the model, and Zach Mainen for discussions. A special thank you to Miguel Bengala for revising and re-plotting all the figures for the chapter under Juan Castineira's and my supervision.

- Chapter 4: Chapter 4 was analyzed by me with the guidance of Alfonso Renart. Miguel Bengala adapted the commitment analysis for the RL agent and flat action bounds agent. We would like to thank Gautam Agarwal, Juan Castiñeiras, Miguel Bengala, and Zach Mainen for discussions.

- Chapter 5: I would like to thank Alfonso Renart, Gautam Agarwal, Juan Castiñeiras, Miguel Bengala, and Zach Mainen for discussions.

This work was funded by a scholarship from the Fundação para a Ciência e a Tecnologia of the Portuguese ministry of education and science (SFRH / BD / 51896 / 2012), and by the Champalimaud Foundation.

Table of contents

Table of contents xi

List of Figures xxxiii

1 Introduction: Decision making in the brain of embodied agents 1
- 1.1 Philosophy and the study of behavior . . . 2
  - 1.1.1 Philosophy of mind . . . 3
- 1.2 Psychophysics: Relating sensation to physical stimuli . . . 4
  - 1.2.1 Perceptual decision making . . . 5
- 1.3 Models of decision making . . . 6
  - 1.3.1 Cognitive models of choice behavior . . . 6
  - 1.3.2 Evidence of computational models in the brain . . . 7
  - 1.3.3 Embodied models of decision making . . . 7
- 1.4 Online readouts of behavior . . . 8
- 1.5 Normative models of decision making . . . 9
- 1.6 Commitment . . . 10
- 1.7 Aim of the thesis . . . 10

2 Online measures of behavior during a decision making task 17
- 2.1 Abstract . . . 18
- 2.2 Introduction . . . 19
- 2.3 Methods of Experiment 1 . . . 20
  - 2.3.1 Subjects . . . 20
  - 2.3.2 Sound Stimulus . . . 21
  - 2.3.3 Online measures of behavior . . . 27
  - 2.3.4 Syncing sounds to other sensors . . . 27

2.4 Results of Experiment 1 . . . 29
2.4.1 Accuracy . . . 29
2.4.2 Reaction Time . . . 29
2.4.3 Choice Trajectory Analysis . . . 30
2.5 Discussion of Experiment 1 . . . 33
2.6 Motivation for Experiment 2 . . . 34
2.7 Methods of Experiment 2 . . . 35
2.7.1 Subjects . . . 35
2.7.2 Task . . . 35
2.7.3 Stimulus . . . 36
2.8 Results of Experiment 2 . . . 38
2.8.1 Accuracy . . . 38
2.8.2 Reaction Time . . . 39
2.8.3 Trajectory Analysis . . . 41
2.8.4 Stimulus Reconstructions . . . 45
2.9 Discussion . . . 51
2.10 Acknowledgments . . . 54

3 Optimal decision making model for temporally extended actions 57
3.1 Abstract . . . 58
3.2 Introduction . . . 58
3.3 Methods . . . 60
3.3.1 Task . . . 60
3.3.2 Partially Observable Markov Decision Processes and Bayesian Belief . . . 60
3.3.3 Bellman equation for embodied 2AFC task . . . 62
3.3.4 Belief Transition Probabilities . . . 65
3.4 Results . . . 74
3.4.1 Normative model of the disembodied 2AFC task . . . 74
3.4.2 Normative model of our embodied 2AFC task, fixing arena width and time limit parameters . . . 79
3.4.3 Decision bounds for the embodied normative agent reflect the agent's commitment to a choice . . . 81
3.4.4 Reliability: Trajectories for repeated presentations of the same stimulus . . . 83
3.4.5 Varying Time Cost with Fixed Arena Width . . . 86

3.4.6 Performance of the embodied vs disembodied agent for fixed arena width and varying time limit parameters 89
3.4.7 Embodied normative agent behavior for varying arena widths and fixed time limit parameters 90
3.4.8 Embodied normative agent behavior for varying arena widths and varying time limit parameters 94
3.4.9 Discussion 96
3.5 Acknowledgments 99

4 A sequential decision making task for studying temporally extended actions 105
4.1 Abstract 106
4.2 Introduction 107
4.3 Methods 107
4.3.1 Task 107
4.3.2 New Stimulus 108
4.3.3 Pilot 2AFC Fixed Duration Experiment 108
4.3.4 Temporally Discounted Reward 109
4.3.5 Subjects and Experimental Procedure 110
4.4 Results 112
4.4.1 Task Stability 112
4.4.2 Accuracy 112
4.4.3 Reaction Times 116
4.4.4 Trajectory Analysis 118
4.5 Discussion 145
4.6 Acknowledgments 147

5 Final Discussion 151
5.1 Final Discussion 151

List of Figures

2.1 We use the interaural level difference (ILD) to create sounds that give the percept of lateralization (left or right). By changing the ILD every 25 milliseconds with controlled probabilities, we create sounds that change their lateralization in time, but are overall louder to one side. The time-varying stimulus envelope is pointwise multiplied by white noise. Top panel: 6 rows of probabilities of sampling from the ILD repertoire determine the mean laterality of the stimulus. Bottom panel: Sampling a sequence of ILDs for medium left difficulty in order to create the stimulus envelope, which is convoluted with white noise in order to produce the stimulus sound waves for each channel. 23

2.2 Example of a 3 second sequence of white noise with time-varying sound pressure level (one channel shown), with the ramp up and ramp down periods of 150 milliseconds calibrated to the subject's auditory mid-line at 60 dB SPL. . . 24

2.3 Top panels show the construction of sound stimuli that are generated by sampling from a random distribution of interaural level differences biased to negative (left) or positive (right) values. Markov chains are used to compute a transition probability matrix that constrains the step size to jump only to neighboring ILD values in order to create a more continuous stimulus that sounds like the sound source is moving inside the subject's head rather than jumping from side to side. Bottom panel shows sample stimulus sequences color coded by difficulty such that Blue denotes stimuli to the right, Red stimuli to the left, and lighter colors denote harder stimuli. 26

2.4 Top panels shows how triggers produced via NI DAQ or through the sound card itself can be used to provide information about the start of the sound on the computer running the task to a separate computer for online measurements. The middle panel shows how digital triggers can be used to inform the separate computer about the end of a sound. Bottom panel shows a 5ms lag between the execution of an early stop to the stimulus and the effective time that we can change the state of the NI DAQ digital out. 28

2.5 Psychometric Curves for subject 1 (left) and subject 2 (right). Post training data is shown for two subjects completing at least 5 sessions of 144 trials. Each session lasted less than 15 minutes. Stimuli were presented in an sound proof booth over calibrated headphones. Colored dots denote the mean proportion of left responses for a particular mean ILD value across all sessions. Error bars show standard deviations computed across sessions. 29

2.6 Response Times for subject 1 (left) and subject 2 (right). ILD Conditions is defined as the bin centers of the histogram of experienced ILDs with 6 equal bin sizes used to compute the psychometric curves of 2.5. The exact values of ILD conditions can vary between the subjects. Response times appear to decrease as the ILD conditions become easier (darker colors), however, the trend is only significant for subject 1's left sided responses. Error bars show standard deviations. 30

2.7 Choice trajectories for subject 1 (top) and subject 2 (bottom). Red shifted colors indicate responses to a left biased stimulus while Blue denotes responses to a right biased stimulus, with lighter colors indicating higher difficulty. 31

2.8 Choice trajectory vs stimulus strength (absolute average ILD across the trial) for subject 1 (left) and subject 2 (right). Blue circles are data. Red line shows linear fit with negative weights for both subjects indicating that the Trajectory Length decreases with increasing mean absolute ILD (decreasing difficulty), $p<0.001$. 32

2.9 Choice trajectory vs stimulus strength (absolute average ILD across the trial) for subject 1 (left) and subject 2 (right). Blue circles are data. For correct trials (left plots), Red line shows linear fit with negative weights for both subjects indicating that the Trajectory Length decreases with increasing mean absolute ILD (decreasing difficulty), $p<0.001$. For incorrect trials (right plots), Blue line shows linear fit with negative weights for both subjects indicating that the Trajectory Length decreases with increasing mean absolute ILD (decreasing difficulty), Not Significant. 33

2.10 Schematic of temporal sequence of events in the task. . . . 36

2.11 Top left: Sample easy left stimulus. Bottom Left: Uniform steady state distribution the stimulus is sampled from. Top right: Left channel white noise stimulus. Bottom right: Right channel white noise stimulus. 37

2.12 Psychometric curve for all subjects computed from 1080 trials post-training. Colored dots denote the mean proportion of left responses for a particular mean ILD value across all sessions. Error bars show standard deviations computed across 12 sessions. 38

2.13 Four panels with 3 rows each show the performance of each subject across sessions. Psychometric curve bias (top), d' (middle) and accuracy (bottom) shows performance remained relatively stable across sessions and days. For the d' plots, the horizontal line marks d'=1. For the Percent Correct plots, the horizontal line marks 70 percent accuracy. Note that subjects 1 and 2 never fall below these performance thresholds, while subjects 3 and 4 were below only 2 and 1 sessions respectively. 39

2.14 Reaction time as a function of the long-term mean of the stimulus (left) and the experienced mean of the stimulus (right) for each subject per row. Blue circles show data, red lines linear fits with significant absolute mean ILD coefficient, and blue lines linear fits that don't have a significant absolute mean ILD coefficient. The plots on the right column show a significant linear decrease in reaction time with increasing absolute experienced mean trial ILD (decreasing difficulty) for all subjects. The decrease in reaction time with decreasing absolute long-term mean of the stimulus is also significant for 3 of the 4 subjects. 40

2.15 Experiment 2 trajectories from each subject's last 150 trials. Each row is one subject. Left 2 columns show the x-axis in time. Right 2 columns show the trajectories in xy. 41

2.16 Clustering of experiment 2 subject 4 data into stereotyped vs hesitant trajectories using k-means with 2 clusters on the trajectory parameters (movement start time, trajectory length, trajectory speed, and acceleration). Note all left ending trajectories were flipped to end on the right prior to clustering for ease of visualization. Left panel shows the trajectories in time, center panel shows them in the xy coordinates of the arena, and the right panel shows the resulting clusters in Length (top) and Speed (bottom) vs Movement Time 42

2.17 We observed more hesitant trajectories for experiment 2 (right panel) than for experiment 1 (left panel). Color code of legend identifies each subject. 43

2.18 For all subjects we observed a significant increase in trial trajectory length with decreasing absolute mean ILD values (increasing difficulty). Blue circles show data, red lines show linear fits with significant coefficient for the independent variable. 43

2.19 Trajectory length analysis for correct and incorrect trials separately. The increase in trajectory length with increasing trial difficulty is significant for subject 1 and subject 3 for both correct and incorrect trials, while subject 2 shows it only for correct trials and subject 4 has a significant effect only for incorrect trials. The results of this analysis are inconclusive regarding confidence effects. Blue circles show data, red (blue) lines show linear fits with significant (not significant) coefficient for the independent variable. 45

2.20 Schematic of training a decoder to predict the moment-to-moment stimulus from a window of eye and hand data. For each stimulus segment to be predicted a window of 500 milliseconds of eye and hand data, which moves forward with the evolving stimulus. Top panel shows the target variable (stimulus), and the corresponding predictor windows (Hand and Eye) for time-step 1. Bottom panel shows the same variables for time-step 2. 47

2.21 Regularized (ridge) linear regression of the moment-to-moment stimulus onto the eye and hand data for all subjects. Regularized (ridge) linear regression of the moment-to-moment stimulus onto the eye and hand data was performed, using for each stimulus segment to be predicted a window of 500 milliseconds of eye and hand data, which moves forward with the evolving stimulus. Post-stimulus kernel refers to the prediction coefficient (β) of each sample in each window of hand and eye data. For each plot, the top panel shows the prediction coefficients kernel fit on 60 percent of the data (training set) for the hand data and the bottom panel eye data. The hand and eye data were z-scored prior to fitting the model. . 49

2.22 Reconstructed stimuli vs. actual stimuli sampled from the test set, not used during training. The title for each plot shows the correlation between the two time-series. Legend: Blue line shows the stimulus, and red line the reconstruction from arm and eye data. 50

2.23 Histograms of the correlation between actual and reconstructed stimuli for all subjects for the test-set. A p-value was calculated for each mean correlation coefficient by shuffling the varying part of the actual stimuli in order to create a distribution of surrogate mean correlations to the reconstructions. Mean correlation for all subjects, except subject 2, was greater than zero and significant. Subject 2 (top right) has the lowest mean correlation, which is also not significant. This is to be expected from her rather ballistic trajectories, faster reaction times, and prediction coefficients kernel which cuts off after 400 milliseconds into the stimulus. 51

3.1 Belief transition probabilities after 5 samples of stimulus (left) and after 15 samples (right). Since the values of belief and probabilities are discrete and not equally spaced, the "squares" in the plot were scaled to *approximately* represent the correct distances in belief space. (the lack of the symmetry that should exist around 0.5 on both axis is a result of *approximating*). Colorbar shows the probability p(g'|g). 71

3.2 New stimulus based on a binomial distribution of ILDs. Black dots represent the stimulus. The green line is an optimal agent's belief that it is experiencing a right-sided stimulus, $p(\pi \geq 0.5)$, given the evidence until that time. Different difficulties can be created by changing the probability, π, of the stimulus being on the right, where π values less (greater) than 0.5 have a left (right) bias. The green line represents the belief resulting from perfect observations (noiseless) and the dashed grey lines correspond to the belief of 5 different trials with the same stimulus, but with noisy observations (stimulus can be "flipped" with probability $\alpha = 0.2$). After noticing that both easy (top) and a hard (bottom) stimuli can be generated with the same sampling probability of being right (0.6), we decided to sample stimuli for the model from this distribution and its left counterpart. 72

3.3 Reaction times of a normative model of the standard 2AFC task with our stimulus. Top panel: Stronger biases to the left (right) are represented by more negative (positive) values on the horizontal axis. Bottom panel: left and right trials are collapsed by taking the absolute value of the difference of the number of right (NR) and left (NL) counts relative to the total number of counts), the stimulus strength (3.23). Colorbar represents the conditional probability of the reaction times given a certain stimulus strength. Black dots represent the mean reaction time for each bin of stimulus strength. Red line is the linear regression of Reaction time onto the stimulus strength. 75

3.4 Reaction time as a function of the absolute experienced stimulus strength. Colorbar represents the conditional probability of the reaction times given a certain experienced stimulus strength. Black dots represent the mean reaction time for each bin of experienced stimulus strength. Red line is the linear regression of Reaction time onto the experienced stimulus strength. 76

3.5 Normative behavior for a single press 2AFC task (arena size 3, meaning left, right and wait states). Left panel: Reaction times increase as the cost of time is decreased (equivalent to increasing the time limit), error bars denote standard deviations. Right panel: Psychometric curves sharpen with decreased cost of time. On both panels, the legend shows the Time Cost converted to the equivalent time limit y (which is proportional to the time at which a positive score is no longer possible). The right panel legend additionally shows the slope parameter for the psychometric curve for each Time Cost. . 76

3.6 Proportion of correct responses increases with decreasing time cost (increasing time limit) whether performance is quantified as a function of the long term mean of the stimulus, p (left panel), or as a function of the experienced stimulus (right panel). On both panels, the legend shows the Time Cost converted to the equivalent time limit y (which is proportional to the time at which a positive score is no longer possible). 77

3.7 Value as a function of the agent's belief. To the left, a slice at t=1 s is shown, representing the values of waiting (black), choosing left (red) or choosing right (blue) for all possible beliefs. To the right, the value function is shown for all times, with larger dots representing the intersection at each time of the curves of waiting with the ones of moving left (large red dots) and the curves of waiting with the ones of moving right (large blue dots). These points correspond to the decision boundaries, where the agent should move left (right) if belief is smaller (bigger), than the one of the decision boundary. Large green dots correspond to the intersection of the value curves for moving left and moving right. 78

3.8 Time evolution of decision boundaries in belief space for different time limits. The decision boundaries collapse in time. The distance between the bounds widens as the cost of time is decreased (time limit, y, is increased). Each pair of mirrored lines represents the threshold or decision bound that the agent's belief must cross at each time in order to make a choice for a particular time limit imposed by a cost for time passing. The legend shows the Time Cost converted to the equivalent time limit y (which is proportional to the time at which a positive score is no longer possible). 79

3.9 Left panel: Number of presses as a function of stimulus strength, for an arena width of 17. The colour scale represents the number of counts and is logarithmic. Black dots represent the mean number of presses for each bin of stimulus strength. Red line is the linear regression of Number of Presses onto Stimulus Strength. Right panel: Reaction times as a function of stimulus strength for an arena width of 17. The colour scale represents the conditional probability of the reaction times given a certain stimulus strength. Black dots represent the mean reaction time for each bin of stimulus strength. Red line is the linear regression of Reaction time onto Stimulus Strength. 80

3.10 Normative agent trajectories for an arena of width 17, back-propagation from 14 seconds, unit cost per action 0.01 with stimulus noise $\alpha = 0.2$ for an easy (left) and a hard (right) trial. Black dots on the borders of each plot show the stimulus, and thin red line the corresponding Bayesian belief. A thick red (blue) line on the left (right) plot shows the decision trajectory of the agent for one presentation of that stimulus. 81

3.11 Value function for all times and for current positions (a)x=-4, (b)x=0, (c)x=4, for an arena width of 17, cost of movement 0.01, cost of time 0.018 and back-propagation from 14 seconds. Small red, blue and black dots represent the values of moving left, moving right and waiting at the current position, respectively. The larger dots represent the intersection at each time of the curves of waiting with the ones of moving left (large red dots) and the curves waiting with the ones of moving right (large blue dots). Large green dots correspond to the intersection of the curves of moving left and moving right. When the agent's belief hits the left (right) bound at the initial position, x = 0, she moves to x = -1 (x=1). Note the value of waiting at x = -4 (x = 4) marked by the green dots is closer to the red (blue) dots denoting the left (right) bound. 83

3.12 Trajectories for 5 repetitions of an easy stimulus and the resulting beliefs (left panels) and trajectories (right panels). Easy stimuli result in identical trajectories for both noiseless (top) and noisy (bottom) integration. Black dots on the borders of each plot show the stimulus. On the left plots, dashed colored lines show the Bayesian beliefs for repeated presentations of the stimulus. Thick colored lines on the right side plot shows the decision trajectories of the agent elicited by the beliefs on the left plots. 84

3.13 Trajectories for 5 repetitions of a hard stimulus and the resulting beliefs (left panels) and trajectories (right panels). For noiseless integration trajectories are identical (top), however with noisy integration (bottom panels) some variability in the output trajectory is introduced the more variable the early part of the stimulus. Black dots on the borders of each plot show the stimulus. On the left plots, dashed colored lines show the Bayesian beliefs for repeated presentations of the stimulus. Thick colored lines on the right side plot shows the decision trajectories of the agent elicited by the beliefs on the left plots. Trajectories remain strikingly qualitatively similar. 85

3.14 Normative agent trajectories for an arena of width 17, time limit 4, unit cost per action 0.01 with stimulus noise $\alpha = 0.2$ for 5 repeats of an easy stimulus (left) and a hard stimulus (right).The red thin lines represent the evolution of belief in the different trials, while blue lines of varying markers denote trajectories that end in the right (figure top) side. Red color likewise denotes trajectories that ended on the left. 86

3.15 Psychometric curves for different costs of time/time limits. The legend shows the Time Cost converted to the equivalent time limit y (which is proportional to the time at which a positive score is no longer possible), and the slope parameter for the psychometric curve for each time limit. 87

3.16 Proportion of correct responses as a function of the time cost. In the left panel, the accuracy of responses is evaluated according to the whole stimulus, while in the right panel it is evaluated according to the section of the stimulus experienced by the agent until the moment of their final choice. The legend shows the Time Cost converted to the equivalent time limit y (which is proportional to the time at which a positive score is no longer possible). 87

3.17 Signatures of the behaviour of the embodied agent for fixed arena width of 17 and $\alpha = 0.2$ and varying time costs/limits. Error bars show standard deviations. The legend shows the Time Cost converted to the equivalent time limit y (which is proportional to the time at which a positive score is no longer possible) 88

3.18 The left (right) panel shows the action bounds to move one step left or right for all times at position x = -1 (x = 1). Color code denotes the bounds for different time limits, for an arena of size x = 17 and stimulus noise parameter $\alpha = 0.2$. As the time limit increases, the bounds get wider. Note that for all time limits, the beliefs necessary to move left (right) indicated by the bottom (top) bound are closer (further) to a fully ambiguous belief of 0.5. Previous movements to the left (right) of the center of the arena makes the left (right) actions more valuable, resulting in a positive feedback loop that biases and eventually commits the agent to the left (right). The legend shows the Time Cost converted to the equivalent time limit y (which is proportional to the time at which a positive score is no longer possible) 89

3.19 Psychometric curves for a disembodied agent (left panel) vs an embodied agent (right panel) for the same costs of time (time limits y). The legend shows the Time Cost converted to the equivalent time limit y (which is proportional to the time at which a positive score is no longer possible), and the slope parameter for the psychometric curve for each time limit. 90

3.20 Performance for our task going from the standard single press 2AFC task (x3) to two press (x5) and larger arenas. Small embodied arenas have similar performance as the standard 2AFC task, with perhaps a small improvement for size (x7) and (x9) but for larger arenas performance drops sharply. The legend shows the arena width, and the slope parameter for the psychometric curve for each width. 91

3.21 Proportion of correct responses as a function of the arena width with a fixed time limit of 4 (time cost = 0.018). In the left panel, the accuracy of responses is evaluated according to the whole stimulus, while in the right panel it is evaluated according to the section of the stimulus experienced by the agent until the moment of their final choice. For both panels, the legend shows the arena width. 91

3.22 Signatures of the behavior of the embodied agents from (Figure 3.20) for fixed time limit of 14 and $\alpha = 0.2$ and varying arena widths. Error bars show standard deviations. For all panels, the legend shows the arena width. 92

3.23 The left (right) panel shows the action bounds to move one step left or right for all times at position x = -1 (x = 1). Color code denotes the bounds for different arena widths, for fixed time limit 14 and stimulus noise parameter $\alpha = 0.2$. As the arena width increases, the bounds collapse earlier while maintaining their width. Note that as observed in (Figure 3.18), for all time limits, the beliefs necessary to move left (right) indicated by the bottom (top) bound are closer (further) to a fully ambiguous belief of 0.5. Previous movements to the left (right) of the center of the arena makes the left (right) actions more valuable, resulting in a positive feedback loop that biases and eventually commits the agent to the left (right). For all panels, the legend shows the arena width. 93

3.24 Signatures of the behavior for a range of arena widths, time limits, and noise parameter $\alpha = 0.2$. The color code is in the scale of each plot's y-limits. 95

4.1 Fixed duration 2AFC experiments. We tested 6 naive subjects (all different from the main experiment subjects), with only the hard conditions, p = 0.6 (Right Stimulus) and p = 0.4 (Left Stimulus), at different fixed durations of 1, 2, 3, 4, 5, 7, 9 , and 12 seconds in separate session blocks of 40 trials per duration with a break in between. As expected, as subjects were allowed to experience shorter stimuli their accuracy decreased 109

4.2 This is a schematic of the expected reward, which grows in time as subjects accumulate more evidence and increase their probability of answering correctly (top), a constant time cost which results in a linearly accumulating point reduction (middle), and the addition of these two quantities resulting in the expected value for completing the task as a function of time (bottom). 110

4.3 Stimulus Repeats. Top panel: Each column shows a unique leftward stimulus and 8 responses from each subject. Each row shows a different subject. Bottom panel: Mirrored rightward stimuli. Black dots on each plot denote the stimulus. Blue (red) colored trajectories end on the right (left). 111

4.4 Scored Points (top row), Percentage correct (middle row), and Reaction times (bottom row) are plotted against session number for all subjects in order to investigate whether any linear trends exist, which would suggest learning is still ongoing. No such trends can be detected at a significance of $\alpha = 0.5$, suggesting performance is stable across sessions. Error bars show standard deviations. 112

4.5 Psychometric curves for 4 subjects evaluating performance according to the long-term mean of the stimulus. Thin black lines show performance for each of the 10 sessions of 72 trials, thick red lines show average psychometric over all 720 trials for each subject. Red to blue graded circles show session performance for 6 difficulty conditions. 113

4.6 Average psychometric curve for 4 subjects evaluating performance according to the long-term mean of the stimulus. Thin black lines show performance for each subject over 720 trials, thick red lines shows the mean for all subjects. Red to blue graded circles show session performance for 6 difficulty conditions. 114

4.7 Psychometric performance as a function of the difficulty of the experienced stimulus 3.23 for all 4 subjects. For each plot, thin gray lines show each session's psychometric, and the thick red line shows the psychometric for all 720 trials. . 115

4.8 Average psychometric curve for 4 subjects evaluating performance according to the mean of the experienced the stimulus. Thin black lines show performance for each subject over 720 trials, thick red lines shows the mean for all subjects. Red to blue graded circles show session performance for 6 difficulty conditions. 116

4.9 Reaction times for 4 subjects decrease as the mean of the experienced stimulus increases (task becomes easier) (3.23). Black dots show mean reaction times for each difficulty bin, red line shows the linear fit of RT onto stimulus difficulty, and the color bar the counts of RT for each bin. 117

4.10 Reaction times decrease as the mean of the experienced stimulus increases (task becomes easier) (3.23). 2280 trials collapsed across all 4 subjects. Black dots show mean reaction times for each difficulty bin, red line shows the linear fit of RT onto stimulus difficulty, and the color bar the counts of RT for each bin. 118

4.11 The number of presses across all 4 subjects increases with increasing difficulty of the experienced stimulus (3.23 $\rightarrow$ 0). Black dots show mean Number of Presses for each difficulty bin, red line shows the linear fit of Number of Presses onto stimulus difficulty, and the color bar the counts of Number of Presses for each bin. 119

4.12 Number of presses increase with increasing difficulty of the experienced stimulus (3.23 $\rightarrow$ 0). 2280 trials collapsed across all 4 subjects. Black dots show mean Number of Presses for each difficulty bin, red line shows the linear fit of Number of Presses onto stimulus difficulty, and the color bar the counts of Number of Presses for each bin. 120

4.13 Each row shows one subject's distributions of inter-press intervals (IPI) for all responses. Red line shows the peak IPI. . 121

4.14 Inter-press-interval (IPI) distributions for all subjects (each column) across 6 different difficulty conditions (each row). Red line marks the peak, magenta line marks the 95th percentile. 122

4.15 Across all 4 subjects inter-press-interval (IPI) decreases with decreasing difficulty of the experienced stimulus (3.23 $\rightarrow$ 1). Black dots show mean IPI for each difficulty bin, red line shows the linear fit of IPI onto stimulus difficulty, and the color bar the counts of IPI for each bin. 123

4.16 Inter-press-interval decreases with decreasing difficulty of the experienced stimulus (3.23 → 1). 2280 trials collapsed across all 4 subjects. Black dots show mean IPI for each difficulty bin, red line shows the linear fit of IPI onto stimulus difficulty, and the color bar the counts of IPI for each bin. 124

4.17 Even after the removal of any IPIs > 0.5 seconds, inter-press-interval decreases with decreasing difficulty of the experienced stimulus (3.23 → 1). 2280 trials collapsed across all 4 subjects. Black dots show mean IPI for each difficulty bin, red line shows the linear fit of IPI onto stimulus difficulty, and the color bar the counts of IPI for each bin. 125

4.18 For 3 of 4 subjects, inter-press-interval decreases with increasing duration in the trial. Black dots show mean IPI for each difficulty bin, red line shows the linear fit of IPI onto stimulus difficulty, and the color bar the counts of IPI for each bin. . 126

4.19 Inter-press-interval decreases with increasing duration in the trials. The effect is very small but significant for 2880 trials across 4 subjects. Black dots show mean IPI for each difficulty bin, red line shows the linear fit of IPI onto stimulus difficulty, and the color bar the counts of IPI for each bin. 126

4.20 Left panel shows 8 responses to repeats of the same stimulus in red, and 8 responses randomly chosen from the 112 responses to the other 14 unique stimuli with the same difficulty condition, p(R), in black (note that these other 112 responses contain 8 repeats of each of the other 14 stimuli with the same condition). Right panel shows the pairwise correlations of the responses shown on the corresponding left panel. Colorbar is the correlation. The matrix is sorted such that the first 8 elements correspond to responses of repeats of the same stimulus, and elements 9-16 show responses to randomly chosen stimuli sampled from the exact same difficulty condition, p(r). 127

4.21 Each left panel shows 8 responses to repeats of the same stimulus in red, and 8 responses randomly chosen from the 112 responses to the other 14 unique stimuli with the same difficulty condition, p(R), in black (note that these other 112 responses contain 8 repeats of each of the other 14 stimuli with the same condition). Each right panel shows the pairwise correlations of the responses shown on the corresponding left panel. Colorbar is the correlation. The matrices are sorted such that the first 8 elements correspond to responses of repeats of the same stimulus, and elements 8-16 show responses to randomly chosen stimuli sampled from the exact same difficulty condition, p(r). 128

4.22 For all 4 subjects, difference between the mean of the same-stimulus correlations sub-matrix and the mean of the same-condition correlations sub-matrix depends on the difficulty of the stimulus. For each unique stimulus matrix of correlations A, we create 2 sub-matrices, one of the correlations between the responses to repeats, Repeats = A(1:8, 18), and one of the correlations between each unique repeat response and the 8 same p(R) condition responses, Same Conditions = A(1:8, 9:16). We eliminate identity and duplicate entries from the Repeats sub-matrix by further sub-sampling only the upper triangular matrix. Finally, we subtract the mean of the Same Conditions sub-matrix from the mean of the Repeats sub-matrix and find that this difference depends on the difficulty of the stimulus. Cyan (magenta) line shows the linear regression of the difference between the mean repeats correlations and the mean same conditions correlation onto the stimulus strength for leftward (rightward) trials. 129

4.23 For each subject, an average correlations matrix is computed by averaging all 90 pairwise correlations matrices comparing responses to stimulus repeats to responses of stimuli sampled from the same condition. Matrices are sorted such that the first 8 elements correspond to responses of repeats of the same stimulus, and elements 9-16 show responses to randomly chosen stimuli sampled from the exact same difficulty condition, p(r). 130

4.24 The difference between the mean of the same-stimulus correlations sub-matrix and the mean of the same-condition correlations sub-matrix depends on the difficulty of the stimulus, data collapsed across 4 subjects. Cyan (magenta) line shows the linear regression of the difference between the mean repeats correlations and the mean same conditions correlation onto the stimulus strength for leftward (rightward) trials. . . 131

4.25 Average correlations matrix is computed by averaging all 90 pairwise correlations matrices comparing responses to stimulus repeats to responses of stimuli sampled from the same condition, collapsed across all 4 subjects. Matrices are sorted such that the first 32 elements correspond to responses of repeats of the same stimulus, and elements 33-64 show responses to randomly chosen stimuli sampled from the exact same difficulty condition, p(r). The difference between the mean of the same-stimulus correlations sub-matrix and the mean of the same-condition correlations sub-matrix is significant for all subjects ($p<0.001$ bootstrapping test). 132

4.26 P(L,W,R|E) at different locations and time intervals in the arena for subject 1. For each row, the leftmost (rightmost) plots represent the leftmost (rightmost) spaces in the arena, and the center plot is the center of the arena. Time is split from 0 to 2 seconds (top row), 2 to 4 seconds (middle row), and 4 to 7 seconds (bottom row). At each time step, we count the number of times the subject waited, moved left, or moved right within the next 0.25 seconds, given how much evidence the subject has experienced. Evidence is defined as 3.3.4.3 over the last N samples, excluding the latest M to account for the delay between experiencing a stimulus and making a deliberate press. Results are shown for $N = 7$ and $M = 3$. . 133

4.27 P(L,W,R|E) at different locations and time intervals in the arena for subject 2. For each row, the leftmost (rightmost) plots represent the leftmost (rightmost) spaces in the arena, and the center plot is the center of the arena. Time is split from 0 to 2 seconds (top row), 2 to 4 seconds (middle row), and 4 to 7 seconds (bottom row). At each time step, we count the number of times the subject waited, moved left, or moved right within the next 0.25 seconds, given how much evidence the subject has experienced. Evidence is defined as 3.3.4.3 over the last N samples, excluding the latest M to account for the delay between experiencing a stimulus and making a deliberate press. Results are shown for N = 7 and M = 3. . 134

4.28 P(L,W,R|E) at different locations and time intervals in the arena for subject 3. For each row, the leftmost (rightmost) plots represent the leftmost (rightmost) spaces in the arena, and the center plot is the center of the arena. Time is split from 0 to 2 seconds (top row), 2 to 4 seconds (middle row), and 4 to 7 seconds (bottom row). At each time step, we count the number of times the subject waited, moved left, or moved right within the next 0.25 seconds, given how much evidence the subject has experienced. Evidence is defined as 3.3.4.3 over the last N samples, excluding the latest M to account for the delay between experiencing a stimulus and making a deliberate press. Results are shown for N = 7 and M = 3. . 135

4.29 P(L,W,R|E) at different locations and time intervals in the arena for subject 4. For each row, the leftmost (rightmost) plots represent the leftmost (rightmost) spaces in the arena, and the center plot is the center of the arena. Time is split from 0 to 2 seconds (top row), 2 to 4 seconds (middle row), and 4 to 7 seconds (bottom row). At each time step, we count the number of times the subject waited, moved left, or moved right within the next 0.25 seconds, given how much evidence the subject has experienced. Evidence is defined as 3.3.4.3 over the last N samples, excluding the latest M to account for the delay between experiencing a stimulus and making a deliberate press. Results are shown for N = 7 and M = 3. . 136

4.30 P(L,W,R|E) at different locations and time intervals in the arena for all subjects collapsed, with a minimum refractory period of 0.2 seconds between unique presses. For each row, the leftmost (rightmost) plots represent the leftmost (rightmost) spaces in the arena, and the center plot is the center of the arena. Time is split from 0 to 2 seconds (top row), 2 to 4 seconds (middle row), and 4 to 7 seconds (bottom row). At each time step, we count the number of times the subject waited, moved left, or moved right within the next 0.25 seconds, given how much evidence the subject has experienced. Evidence is defined as 3.3.4.3 over the last N samples, excluding the latest M to account for the delay between experiencing a stimulus and making a deliberate press. Results are shown for N = 7 and M = 3. 139

4.31 P(L,W,R|E) at different locations and time intervals in the arena for 30k simulated trials of the optimal RL agent. For each row, the leftmost (rightmost) plots represent the leftmost (rightmost) spaces in the arena, and the center plot is the center of the arena. Time is split from 0 to 2 seconds (top row), 2 to 4 seconds (middle row), and 4 to 7 seconds (bottom row). At each time step, we count the number of times the subject waited, moved left, or moved right within the next 0.25 seconds, given how much evidence the subject has experienced. Evidence is defined as 3.3.4.3 over the last N samples, excluding the latest M to account for the delay between experiencing a stimulus and making a deliberate press. Results are shown for N = 4 and M = 0, since the agent has perfect memory. 141

4.32 P(L,W,R|E) at different locations and time intervals in the arena for 30k simulated trials of an agent with wide flat action bounds that do not change in space nor time. For each row, the leftmost (rightmost) plots represent the leftmost (rightmost) spaces in the arena, and the center plot is the center of the arena. Time is split from 0 to 2 seconds (top row), 2 to 4 seconds (middle row), and 4 to 7 seconds (bottom row). At each time step, we count the number of times the subject waited, moved left, or moved right within the next 0.25 seconds, given how much evidence the subject has experienced. Evidence is defined as 3.3.4.3 over the last N samples, excluding the latest M to account for the delay between experiencing a stimulus and making a deliberate press. Results are shown for N = 4 and M = 0, since the agent has perfect memory. We set a window of 4 samples to match the window used for the subjects, however setting N to the maximum value of 28 does not qualitatively change the plots. We simulated 30,000 trials with bound height comparable to the bounds observed by the optimal agent of (Figure 4.31) at the 3rd time-step in the center of the arena. 143

4.33 P(L,W,R|E) at different locations and time intervals in the arena for 30k simulated trials of an agent with narrow flat action bounds that do not change in space nor time. For each row, the leftmost (rightmost) plots represent the leftmost (rightmost) spaces in the arena, and the center plot is the center of the arena. Time is split from 0 to 2 seconds (top row), 2 to 4 seconds (middle row), and 4 to 7 seconds (bottom row). At each time step, we count the number of times the subject waited, moved left, or moved right within the next 0.25 seconds, given how much evidence the subject has experienced. Evidence is defined as 3.3.4.3 over the last N samples, excluding the latest M to account for the delay between experiencing a stimulus and making a deliberate press. Results are shown for N = 4 and M = 0, since the agent has perfect memory. We set a window of 4 samples to match the window used for the subjects, however setting N to the maximum value of 28 does not qualitatively change the plots. We simulated 30,000 trials with bound height comparable to the bounds observed by the optimal agent of (Figure 4.31) at the 18th time-step in the center of the arena. 144

Chapter 1

Introduction: Decision making in the brain of embodied agents

HIGHLIGHTS

- In this first chapter we do a literature review concerning the study of behavior, and introduce the field of psychophysics.
- We review the literature on role of the brain in mediating perception and decision-making behavior.
- We present a summary on the use of computational models in the field of decision making.
- We introduce the literature on models of embodied decision making and begin to argue that studying the motor component of a decision might lead to more accurate models of the types of problems facing animals making naturalistic choices.
- We motivate the need for a novel task to study online behavior during the decision period, and summarize the literature on the topic.
- We outline the aims of this thesis and provide a summary of each chapter.

Contents

1.1 Philosophy and the study of behavior

How should we behave? Some of the most inquisitive minds throughout history have attempted to ground their philosophical inquiries into human behavior on pure fictions, such as the presumed desires of one or many Gods. Such fictions remained unchallenged until the invention of scientific thinking, which provides a method for testing hypotheses. Major discoveries achieved through scientific thinking have compelled modern philosophers to integrate into their theories of behavior knowledge from various fields such as logic, sociology, and more recently evolutionary biology and neurobiology. Scientific thinking is rooted in a tradition of skepticism owing greatly to Socrates and his student Plato, founder of the Academy of Athens in ca. 387 BC. Plato lectured on the nature and purpose of all things, including man, relying on an ideal heavenly construct where the perfect Forms of all things exist. According to Plato, these perfect Forms project their perfect qualities to imperfect images or shadows that compose the reality we live in. Every man is an imperfect physical shadow which borrows his qualities from the perfect qualities of the true, perfect man that exist as an idea in the realm of Forms. Our senses are able to directly experience only the imperfect shadow reality, but inferences about the perfect Forms can be made by studying the qualities of the imperfect world through repeated sensory observations and philosophical inquiry. However wrong Plato's premises about the realm of Forms were, his conclusions on the importance of repeated sampling and philosophical reasoning changed the course of history. Plato's students included Aristotle, who - among many other seminal achievements - described a general framework for constructing logical arguments. The remarkable feat of Aristotle's logical argument is that if one's premises are right, and deduction is applied thereafter, one should arrive at true conclusions every time. Aristotle's famous argument, "Socrates is a man. All men are mortal. Therefore, Socrates is mortal" demonstrates the basic form of what he called a "perfect" syllogism, because no further premises are needed in order to make the conclusion follow the premise. This systematic mode of logical inference

eventually evolved into our modern recipe for rejecting false hypotheses in order to discovery the truth, the scientific method [Garvey, 2012].

1.1.1 Philosophy of mind

In the first half of the 15th century, René Descartes wrote his Meditations on First Philosophy, in which he remarks that that nearly everything he believes has come to him from his senses. However, soon after this assertion comes his most famous phrase, "Having thought carefully about it, and having scrupulously examined everything, one must then, in conclusion, take as assured that the proposition, I am, I exist, is necessarily true, every time I express it or conceive it in my mind" [Descartes, 1998] in which Descartes discards the senses on account that they sometimes deceive us, for example when we dream and sense things that are not real. Descartes' Meditations did much more than proving his own existence, or at least the existence of his thoughts, his proposition that all knowledge enters the brain through the senses set the stage for empirically based knowledge. Unlike Plato's perfect Forms, the senses can be investigated through empirical observation. Since Descartes, other philosophers like Voltaire, Locke and Hume have given the senses and the perceptions and the subjective experiences they convey a center stage in their attempts to provide a foundation of knowledge. David Hume in fact boiled down the very identify of a person to perception alone when he wrote, "For my part, when I enter most intimately into what I call myself, I always stumble on some particular perception or other, of heat or cold, light or shade, love or hatred, pain or pleasure. I never catch myself at any time without a perception, and never can observe anything but the perception" [Hume, 1817]. He concludes that any idea of a self is a philosophical delusion. Rather than make a case for or against David Hume's views on perception as the centerpiece of reality, this thesis assumes the importance of studying sensory perceptions, and focuses in particular on the investigation of decisions made upon them.

1.2 Psychophysics: Relating sensation to physical stimuli

Studies on the human mind continued to take the form of philosophy, attempting to tackle consciousness and the fabric of reality itself long after Hume. An argument could be made that all of that philosophy and even modern science doesn't have much to add to Hume's description of consciousness as sequences of perceptions. However, a major breakthrough in the study of behavior and the mind itself occurred when rather than describing qualitatively exactly what consciousness is, and attempting to apply logic to those descriptions, the field of psychophysics began to quantitatively relate the subjective experience of a sensation to the physical stimulus that cause it. The earliest examples of this type of study date back to Ernst Heinrich Weber's 1830s experiments on sensation and touch [Weber, 1834]. Weber was able to compute the minimum change in stimulus intensity between two distinct stimuli required to elicit two physically discernible percepts (the just noticeable difference, jnd). Weber's results inspired Gustav Theodor Fechner to develop new methods to measure the jnd across several modalities, resulting in what he called the Weber-Fechner law, "In order that the intensity of a sensation may increase in arithmetical progression, the stimulus must increase in geometrical progression" [Fechner, 2007]. This relationship can be written mathematically as, $S = K / \ln I$, where I is the intensity of the sensation, S is the strength of the stimulus and K is a constant that depends on the type and sensory modality of the stimulus. The Fechner-Weber law provides near perfect approximations of the relationship between stimulus and sensation, with a few exceptions, most notably when stimulus energy is so small that it is at the threshold of being consciously experienced. Fechner's Element de Psychophysik was published in 1860, marking the birth of the new field of psychophysics [Fechner, 2007]. A key concept in psychophysics is in fact that of a sensory threshold. A sensory threshold is defined as the "smallest amount of stimulus energy required to produce a sensation" [Gescheider et al., 1997]. In general, an organism's sensitivity to external stimuli varies slightly from moment to moment, but experimenters are able to accurately estimate the absolute sensory threshold by averaging repeated measurements of the threshold value [Gescheider et al., 1997].

1.2.1 Perceptual decision making

The logarithmic mental line described by the Fechner-Weber law manifests itself in a subject's probability of response curve to different stimulus intensities, which goes from 0.5 probability of detection for below threshold stimulus values, and grows logarithmically in probability until saturating at 1 for high enough stimulus intensities. Such a quantification of choice behavior is called a psychometric curve, and it is generally observed to have a sigmoidal shape. The 1940s and 50s saw the rise of the a new revolution in behavior research led by a group of individuals that called themselves behaviorists. Ever since, a large focus of study for behaviorists has been describing the shape of these sigmoids to different stimuli across all sensory modalities, as well as the effect of stimulus perturbations on the parameters of the sigmoid. The advances in technology made possible by electricity allowed behavior researchers to present stimuli and record data with unprecedented accuracy. Today, modern technology allows neuroscientists to compute a psychometric curve for neuronal responses, the neurometric curve. A major milestone for psychophysics and the study of perceptual decisions was accomplished by perhaps the most renown behaviorist, B.F. Skinner. During his graduate studies, B.F. Skinner created the operant conditioning chamber to electronically present and control stimuli such as light flashes or speaker produced sounds. Similarly, subject responses are electronically measured and recorded, creating an objective record of the outcome of an experiment. This apparatus is now commonly known as the Skinner box. Skinner published his findings in 1938, revolutionizing the study of behavior [Skinner, 1938]. Neuroscientists and behavioral economists today continue to make great use of the Skinner box in order to obtain the precise timing of electronically measured choices. A Skinner box with two response levers or buttons has been a favorite tool to measure and describe human and animal psychophysical responses. The simple experimental design makes it possible to link choice behavior, for example reaction time distributions and psychometric curves, to formal mathematical models of the underlying cognitive operation required to complete the task.

1.3 Models of decision making

1.3.1 Cognitive models of choice behavior

The data generated during a standard two choice task allows researchers to create descriptive models of the decision making process. These models can express the computation supporting the decision in terms of stimulus signal and noise at the level of the sensory organs and neural decision centers. The drift diffusion model (DDM) is perhaps the most common of such theoretical approaches to the problem of deciding among two choices based on uncertain evidence [Ratcliff, 1978, Ratcliff, 1985, Ratcliff, 2002]. One of the remarkable feats of a standard drift diffusion model is its ability to make predictions about the psychometric curve of a subject by using only the reaction time data to fit the model and vice-versa [Ratcliff, 2002]. Drift diffusion is a name applied to a family of models that describes the deliberation in choice behavior as a "Drunken" random walk taken by a particle to one of two choice boundaries. The mean direction of the particle is a parameter that depends on the strength of the stimulus, and random perturbations to the mean walk are caused by a parameter(s) that captures noise in the sensory observations, noise in the processing of these observations at the decision centers in the brain, or both. The drift diffusion family of models has been found to find equivalent solutions to the sequential probability ratio test [Wald and Wolfowitz, 1948], which aims to provide sequential hypothesis testing [Stone, 1960, Ratcliff et al., 2016]. While the technique was developed for engineering and manufacturing applications, this method has been applied to describe the stopping criterion for human decision processes as early as 1960 [Stone, 1960]. The recent availability of cheap and fast computers to fit the model parameters has popularized the use of drift diffusion models in the field of decision making to for example, characterize stopping criterion (e.g. reaction time and accuracy) [Ratcliff, 1985]. The drift diffusion model has also been used to explain valued based choices [Milosavljevic et al., 2010], and other neural processes such as memory [Ratcliff, 1978]. A recent study modeling normative solutions to the standard 2AFC decision task found that the optimal trade-off between certainty against the cost of time computed using the partially observable Markov decision process framework is equivalent to drift diffusion with time collapsing bounds [Drugowitsch et al., 2012]. That study ex-

tended the classic SPRT result to cases where the reliability of the evidence (what we call 'difficulty') is unknown to the agent [Drugowitsch et al., 2012].

1.3.2 Evidence of computational models in the brain

The big unanswered question until the beginning of this century was whether and how the components of behavior captured by the drift diffusion framework are reflected in the firing rates of neurons in the brain. In order to experimentally address this question, it became necessary to record neural activity while an animal performed a decision task. A pioneer study managed to obtain single cell recordings from the lateral intraparietal cortex of behaving rhesus monkeys. It was found that choice movements were only initiated if neural activity reached a certain threshold, and that stochastic variability in the rate at which neuronal activity grew to that threshold resulted in the distribution of behavioral reaction times [Hanes and Schall, 1996]. Shadlen and colleagues (e.g. [Gold and Shadlen, 2001, Mazurek et al., 2003]) have made significant efforts to advance both experimental and theoretical research regarding evidence accumulation processes in the brain. Recent work by Carlos Brody's group has also observed evidence accumulation recording from neurons in posterior parietal cortex and FOF from rats performing a perceptual decision-making task [Hanks et al., 2015]. Although these findings are consistent with models of evidence accumulation, there are competing views, for example, one which argues that the increasing firing rates observed in the monkey and rat brains reflect an urgency signal rather than evidence accumulation [Cisek et al., 2009].

1.3.3 Embodied models of decision making

The models described above can be considered "cognitive" models since they are only concerned with the final outcome of the task, assuming that the execution of a response movement is handled by downstream motor centers that are independent from and do not affect the computations at the decision centers. While such models manage to capture the psychometric curve and response time distributions of human and animal subjects, the assumption of independence between the decision and motor centers is made out the necessity to simplify the computational problem rather than a belief that this is how biological organisms arrive at naturalistic choices. More

complex computations, for example a distributed representation of choices that takes into consideration biological constraints, such as biomechanical costs for moving, have been proposed [Thura et al., 2012]. Recent studies that model embodied decisions modify the drift diffusion framework to allow for processing of the stimulus in parallel with action initiation and execution, with the basic premise that these processes can bias the decision of the animal in ways that optimise the ecological choices in naturalistic environments [Resulaj et al., 2009, Martinez-Hernandez et al., 2015]. Relative to the purely cognitive models, the embodied models models can result in faster actions when the task at hand is easy, and action corrections when the task is hard [Resulaj et al., 2009, Martinez-Hernandez et al., 2015].

1.4 Online readouts of behavior

As a subject makes a choice, we hypothesized that her online behavior during the deliberation period might reveal hints of the process happening in her decision centers. If a random walk to one of two choices is indeed happening in the subject's brain, can it be observed directly from behavior while it is happening? We found the traditional binary choice task with its single button press response too constraining to allow the study of situations in which online behavior could couple with the mental computations leading to a choice. We were inspired by previous studies replacing button presses with reach trajectories to one of two targets, for example using a computer mouse [Spivey et al., 2005] or a haptic joystick like controller [Resulaj et al., 2009]. These studies observed hesitant trajectories when the choice was difficult, the first study in the form of down the middle trajectories [Spivey et al., 2005], and the second observed occasional changes of mind [Resulaj et al., 2009], which the authors could predict using a two-stage drift diffusion model. However, these studies did not link the moment to moment accumulation process to online movements that could reflect sequences of sub-decisions. We hypothesised that either the simple nature of the stimulus [Spivey et al., 2005] or the extinction of stimulus presentation time during the movement period [Resulaj et al., 2009], prevented these studies from further linking the online evolution of the mental process to online movements. A subsequent study managed to link the mental evidence accumulation to a sub-threshold motor signal prior to the start of choice

movements [Selen et al., 2012]. Our initial objective was to link the process of evidence accumulation to online movements, somehow. After successfully designing an experiment that allowed us to observe movement during the deliberation period, understanding and modeling the conditions that allow such coupling of mental and motor processes to occur became the second objective of this thesis.

1.5 Normative models of decision making

With our online readouts of behavior, we managed to collect data of subjects showing movements during the decision period, and furthermore could show that these movements were predictive of the fluctuations in time of the random stimulus. We knew that the movements reflected a form of hesitation or lack of commitment to either choice, since they appeared mostly during hard trials. The uncommitted or hesitant "down the middle" aspect of our trajectories had previously been reported by [Dale et al., 2007], and corrective movements correlated to fluctuations in the stimulus were previously shown by [Resulaj et al., 2009]. However, the ability to measure the continuous modulation of sequences of behavioral output tuned to a stimulus during a binary choice discrimination constituted a novel approach to the laboratory study of the decision making problem. We next looked for a theory that could explain this phenomenology. The insights from the accumulation of evidence to a bound models were not directly applicable to our data, since those tasks typically ended when the initial choice was made. The changes of mind extension to the drift diffusion framework also did not seem to provide a sufficient framework for our data, which sometimes contained what appeared to be multiple changes of mind. We finally turned to reinforcement learning (RL) to study normative solutions to our evidence accumulation task. We started from a recently published theoretical framework that uses insights from the SPRT and reinforcement learning (RL) to find optimal solutions to the binary discrimination accumulation task, given a temporal constraint [Moreno-Bote, 2010, Drugowitsch et al., 2011, Drugowitsch et al., 2012]. We generalized this framework to include the spatial navigation component in our task by including a spatial constraint, a cost for moving towards the targets in a discrete grid world. Through our simulations, we came to understand the set of circumstances required to elicit

hesitant choice behavior (as well as fully committed behavior). By assuming parameters regarding the signal-to-noise in the perception of the stimulus, as well as parameters regarding the agent's urgency to act, and the amount of motor effort involved in her actions, our model's choice trajectories capture the main phenomenology that we observed in our psychophysical work. Namely, that difficult stimuli cause subjects and agents to make correcting movements, resulting in longer choice trajectories. In the RL agent, this occurs because when uncertainty is high, the agent finds intermediate positions more valuable, as extreme positions make it difficult to correct initial wrong choices. Only longer oscillations in the stimulus cause beliefs extreme enough to act upon. This results in a loose tracking of the stimulus by the agent.

1.6 Commitment

The agent moves when her believes about the stimulus cross belief thresholds. If the agent still has enough time to reach either target, changes of mind occur any time the opposite threshold is crossed. We call this a soft decision bound since the actions are reversible. However, when the agent is closer to a target, it will take her longer to travel back to the other target should she change her mind. This constraint together with the time pressure eventually causes the decision bounds to be final or irreversible, as no other action than approaching the closest target is estimated to ever have higher value at any future time. Thus, the agent must commit to her current choice. We call this a hard or irreversible decision bound. Our framework shows that positive feedback naturally arises when temporally-extended decisions are made by embodied agents, and that this form of "commitment" is likely to generally contribute to the categorical nature of decision-making in naturalistic scenarios facing organisms with temporal and spatial costs.

1.7 Aim of the thesis

Our goals were to design a decision making task that allowed subjects to move during the deliberation of a binary choice in order to observe or infer an online decision variable directly from behavior. However, we needed to be careful not to depart too much from the classical binary choice configura-

tion, which has the strength that it constrains behavior so that meaningful statistics can be computed from the generated data. This of course proved to be more difficult than we anticipated, but resulted in the collection of novel data, the hesitant decision trajectories, for which providing an explanation became at least as daunting of a task. The explicit questions that are investigated in this thesis are the following:

- How can the evidence accumulation process be studied from the online behavior of human subjects?
- Are online choice trajectories measured during the performance of a binary choice task reflective of the evidence accumulation process?
- Why do subjects display complex decision trajectories when the stimulus is hard?
- What are the circumstances for which complex decision trajectories arise?

Chapter 2 explains the design and construction of Experiment 1, a novel psychophysical task that allowed us to capture movements during the execution of the task. The second half of the chapter is concerned with Experiment 2, which had the aim of creating experimental conditions that allowed us to observe more complex trajectories per experimental session. The modifications implemented in Experiment 2 mark the start of our quest to understand why the complex trajectories arise to begin with. The chapter includes a deeper analysis of our online measures of behavior.

Chapter 3 is concerned with optimal decision making models that can solve a simplified version of our task. This work helped us understand the circumstances for which complex decision trajectories arise, and allowed us to construct a psychophysical task that provides better control the spatial constraints and time urgency with respect to Experiment 1 and 2.

Chapter 4 describes Experiment 3, a psychophysical task with simpler controls than our previous experiments, and more explicit feedback regarding rewards for correct answers against a penalty for time passing. Precise inter-press intervals across difficulties for all 4 tested subjects suggest subjects are performing the same task, and despite this stereotyped mode of

pressing, a distinction between confident vs hesitant decision trajectories in response to the stimulus can be appreciated.

Chapter 5 summarizes and discusses all the above results, and in particular the implications of our findings for binary choices vs more complex designs in Perceptual DM, the interaction between action and belief during sequential sampling problems, cost of action in (perceptual) DM and "geometric" costs, and the nature of commitment.

Bibliography

[Cisek et al., 2009] Cisek, P., Puskas, G. A., and El-Murr, S. (2009). Decisions in changing conditions: The urgency-gating model. *Journal of Neuroscience*. (Cited on page 7.)

[Dale et al., 2007] Dale, R., Kehoe, C., and Spivey, M. J. (2007). Graded motor responses in the time course of categorizing atypical exemplars. *Memory and Cognition*. (Cited on page 9.)

[Descartes, 1998] Descartes, R. (1998). Discourse on Method and Meditations on First Philosophy. (Cited on page 3.)

[Drugowitsch et al., 2011] Drugowitsch, J., Moreno-Bote, R., and Pouget, A. (2011). Optimal decision bounds for probabilistic population codes and time varying evidence. *Nature Precedings*. (Cited on pages 9 and 59.)

[Drugowitsch et al., 2012] Drugowitsch, J., Moreno-Bote, R. N., Churchland, A. K., Shadlen, M. N., and Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. *Journal of Neuroscience*. (Cited on pages 6, 7, 9, 53, 58, 59, 60, 62, 65, 66, 74, 76, 78, 97, 98, 106, 153, and 154.)

[Fechner, 2007] Fechner, G. T. (2007). Elements of psychophysics, 1860. In *Readings in the history of psychology*. (Cited on page 4.)

[Garvey, 2012] Garvey, J. (2012). *The Story of Philosophy: A History of Western Thought*. Quercus. (Cited on page 3.)

[Gescheider et al., 1997] Gescheider, G. A., Thorpe, J. M., Goodarz, J., and Bolanowski, S. J. (1997). The effects of skin temperature on the detection and discrimination of tactile stimulation. *Somatosensory and Motor Research*. (Cited on page 4.)

[Gold and Shadlen, 2001] Gold, J. I. and Shadlen, M. N. (2001). Neural computations that underlie decisions about sensory stimuli. (Cited on page 7.)

[Hanes and Schall, 1996] Hanes, D. P. and Schall, J. D. (1996). Neural control of voluntary movement initiation. *Science*. (Cited on page 7.)

[Hanks et al., 2015] Hanks, T. D., Kopec, C. D., Brunton, B. W., Duan, C. A., Erlich, J. C., and Brody, C. D. (2015). Distinct relationships of parietal and prefrontal cortices to evidence accumulation. *Nature*. (Cited on pages 7, 152, and 155.)

[Hume, 1817] Hume, D. (1817). *A treatise of human nature: being. An attempt to introduce the experimental method of reasoning into moral subjects*. (Cited on page 3.)

[Martinez-Hernandez et al., 2015] Martinez-Hernandez, U., Lepora, N. F., and Prescott, T. J. (2015). Active haptic shape recognition by intrinsic motivation with a robot hand. In *IEEE World Haptics Conference, WHC 2015*. (Cited on page 8.)

[Mazurek et al., 2003] Mazurek, M. E., Roitman, J. D., Ditterich, J., and Shadlen, M. N. (2003). A Role for Neural Integrators in Perceptual Decision Making. (Cited on page 7.)

[Milosavljevic et al., 2010] Milosavljevic, M., Malmaud, J., Huth, A., Koch, C., and Rangel, A. (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*. (Cited on page 6.)

[Moreno-Bote, 2010] Moreno-Bote, R. (2010). Decision confidence and uncertainty in diffusion models with partially correlated neuronal integrators. (Cited on pages 9, 59, 60, and 153.)

[Ratcliff, 1978] Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*. (Cited on pages 6 and 97.)

[Ratcliff, 1985] Ratcliff, R. (1985). Theoretical Interpretations of the Speed and Accuracy of Positive and Negative Responses. *Psychological Review*. (Cited on page 6.)

[Ratcliff, 2002] Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin and Review*. (Cited on page 6.)

[Ratcliff et al., 2016] Ratcliff, R., Smith, P. L., Brown, S. D., and McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in cognitive sciences*, 20(4):260–281. (Cited on pages 6 and 59.)

[Resulaj et al., 2009] Resulaj, A., Kiani, R., Wolpert, D. M., and Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature*. (Cited on pages 8, 9, 51, 52, 145, 151, and 152.)

[Selen et al., 2012] Selen, L. P., Shadlen, M. N., and Wolpert, D. M. (2012). Deliberation in the motor system: Reflex gains track evolving evidence leading to a decision. *Journal of Neuroscience*. (Cited on pages 9, 20, 27, 34, 52, 145, and 151.)

[Skinner, 1938] Skinner, B. F. (1938). The Behaviour of Organisms. *American Psycologist*. (Cited on page 5.)

[Spivey et al., 2005] Spivey, M. J., Grosjean, M., and Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences of the United States of America*. (Cited on pages 8, 20, 27, 34, 51, 52, 145, 151, and 152.)

[Stone, 1960] Stone, M. (1960). Models for choice-reaction time. *Psychometrika*. (Cited on pages 6 and 59.)

[Thura et al., 2012] Thura, D., Beauregard-Racine, J., Fradet, C. W., and Cisek, P. (2012). Decision making by urgency gating: Theory and experimental support. *Journal of Neurophysiology*. (Cited on page 8.)

[Wald and Wolfowitz, 1948] Wald, A. and Wolfowitz, J. (1948). Optimum Character of the Sequential Probability Ratio Test. *Ann. Math. Statist.*, 19(3):326–339. (Cited on pages 6 and 59.)

[Weber, 1834] Weber, E. H. (1834). De Pulsu, resorptione, auditu et tactu. Annotationes anatomicae et physiologicae ... (Cited on page 4.)

Chapter 2

Online measures of behavior during a decision making task

HIGHLIGHTS

- In this chapter we describe a novel psychophysics task in order to measure behavior during the deliberation period of a binary choice task.
- Hesitant trajectories can be observed in a small fraction of the trials in Experiment 1. On average the length of the trajectories increases with increasing stimulus difficulty.
- In order to capture more hesitant trajectories per session, Experiment 2 imposes a time limit to initiate movements from the start of the stimulus.
- In Experiment 2, trajectory length increases significantly with increasing stimulus difficulty for all subjects. No consistent difference can be seen for correct vs. incorrect trials across subjects.
- It is possible to reconstruct the noisy stimulus from linear combinations of eye and hand trajectories. This shows that the online trajectories hold information about temporal features of the stimulus.

Contents

2.1 Abstract 18
2.2 Introduction 19
2.3 Methods of Experiment 1 20
2.4 Results of Experiment 1 29
2.5 Discussion of Experiment 1 33
2.6 Motivation for Experiment 2 34
2.7 Methods of Experiment 2 35
2.8 Results of Experiment 2 38
2.9 Discussion 51
2.10 Acknowledgments 54

2.1 Abstract

The behavioral evidence typically used to infer the mechanisms underlying decision making processes consist of reaction time (RT) and choice, which reflect only the final output of the decision, and is therefore only indirectly informative on the decision process itself. In order to gain more direct access to the decision process from behavior, we set-up an auditory two-alternative forced choice task (2AFC) perceptual decision-making task with two online behavioral read-outs: motor output and pupil size. Changes in pupil size have been associated with altered vigilance [Rajkowski et al., 1994]. For example in one study using chair-restrained, unanesthetized cynomolgus monkeys, pupil dilation was correlated to locus coeruleus neuronal activity, which was higher during behavioral agitation than during goal-directed task behavior [Rajkowski et al., 1994]. In addition to these changes in tonic activity, LC neurons were also phasically responsive to certain sensory stimuli. It may be more correct to think about pupil measures as physiology rather than behavior. In the task, human subjects have to decide whether the overall lateralization of a time-varying white-noise stimulus is left or right. Lateralization is controlled using the interaural level difference (ILD) of the stimulus, delivered through headphones. The instantaneous ILD is chosen

with a given probability every 25 ms from a set of ILD values, and the overall difficulty is controlled by rigidly moving this probability function. Subjects report their choices by moving a joystick-controlled dot in a computer screen to either of two lateral locations and are free to move the joystick throughout the stimulus presentation. Pupil size is monitored through the experiment. Stable psychometric functions confirm that the subjects can reliably perform this task. A small fraction of subject responses (15 %) elicited during the presentation of difficult stimuli are not direct to either target, but rather wiggling between the two targets down the middle. We think subjects' hesitation in the deliberation period is coupled to the stimulus, and therefore to the underlying decision making process, but more data was needed for meaningful analysis. In a second experiment we managed to increase the number of hesitant trajectories displayed by subjects by introducing time pressure on movement onset. Analysis of these trajectories allowed us to establish a correlation between online behavior and the stochastic stimulus.

2.2 Introduction

The study of human and animal choice behavior has traditionally been restricted to laboratory preparations such as the two-alternative forced choice task (2AFC). In a typical 2AFC task, two stimuli are presented successively or side-by-side, and subjects are asked to choose and report one of the two options contingent on their percept of some stimulus feature (e.g. presence, amount, identity), or simply based on their preference for one option, by pressing one of two levers. The two options can be presented simultaneously or in two sequential intervals. The studies on perceptual reports of a stimulus feature belong to the subfield of perceptual decision making, while choices of preference fall under the umbrella of value-based decision making. While we focus on the perceptual decision making literature, significant cross-talk between the subfields means that the state-of-the art techniques are not exclusive to either field. The 2AFC task is said to be unbiased by subjective internal criteria unrelated to the task at hand [Pelli and Farell, 1996]. However, the study of choice in maximally reduced laboratory settings has also precluded these studies from learning about the decision process while the decision is ongoing. For example, Spivey and colleagues showed compelling evidence for continuous uptake of sensory input and dynamic compe-

tition between simultaneously active representations in the performance of a spoken-language task by recording the streaming x, y coordinates of continuous goal-directed hand movement [Spivey et al., 2005]. In that speech recognition study, two pictures appeared on the top left and top right of a computer screen, and subjects were required to make a reach movement from to bottom of the screen towards the correct picture, via computer mouse, according to the word spoken over headphones. Spivey observed that the movement trajectories revealed whether the two choices are phonetic competitors or not, for example candy vs candle compete, and candy vs horse do not compete. We wanted to conduct psychophysical experiments in this fashion, but using perceptual stimuli such as the noisy stimulus utilized by Shadlen and colleagues in [Selen et al., 2012]. We favored such a stimulus because we though that observing changes in behavior as a function of changes in the stimulus parameters would allow us to understand something about the cognitive process in between. Shadlen and colleagues were able to model and predict the changes of mind by extending their bounded accumulation of evidence model to one with two bounds. However, in order to fit their model to human data, the experimenters constrained the evidence accumulation to immediately after the initiation of the reach movement. They extinguished the stimulus as soon as subjects began their choice trajectory. However, we are interested in monitoring behavior while the stimulus is ongoing. Unlike [Selen et al., 2012], we conduct our study in the auditory domain given the relative ease of producing calibrated sounds.

2.3 Methods of Experiment 1

2.3.1 Subjects

Two naive subjects learned to discriminate the stochastic sound stimulus, and by their 3rd session produced reasonable psychometric curves. The subject performed 720 and 1080 trials across 8 and 12 sessions respectively. Two subjects were discarded for having too extreme biases during the calibration routine (a simpler version of this experiment with a single ILD white noise burst per trial).

2.3.2 Sound Stimulus

Below we describe the construction of an auditory stimuli created from sequences of white noise biased to the left or to the right. Our first attempt at creating such stochastic sequences of a controlled stimulus was only used in Experiment 1, but we do not show that data since were more successful in capturing movement during the decision period using the second stimulus. The second stimulus is also stochastically sampled from a biased ILD distribution, but it moves slower through the subject's sound space, this is accomplished by constraining the maximum ILD difference from one sample to the next using Markov Chains. We only show the analysis of Experiment 1 for the second stimulus, and this second stimulus was the only one used in Experiment 2. In chapter 3 we replace the stimulus by much simpler sequences of white noise, where each individual chunk of stimulus is a fully lateralized burst with a binomial probability of being to the left or to the right.

We induce the percept of a lateral sound by manipulating the difference in the intensity of sound arriving at each ear, the interaural level difference (ILD), which is a major cue assisting sound localization in mammals. We play gaussian noise to both ears while keeping the average binaural sound pressure level (SPL) constant, and apply some ILD to the baseline noise from a finite set of possible ILDs every few milliseconds. By controlling the probability of drawing each ILD, we create stimuli with different levels of average lateralization for which the subject must indicate which side was on average "louder." We present sounds over calibrated headphones (Sennheiser HD 280 Pro) using an internal sound card (Asus Xonar Essence ST). In order to make our stimuli salient but comfortable to listen to during extensive testing, we fixed the binaural level at 50 dB SPL. Our experience is that a 3 dB ILD is necessary to attain 90% performance in naive subjects, and invariable perfect performance is attained with sounds lateralized to either side by 6 dB ILD. To reduce the ambiguity of our time-varying stimulus, we decided to use sounds lateralized by more than 4 dB to either side. Our time-varying stimulus is constructed with the repertoire of ILDs [-12 -8 -4 0 4 8 12] dB, where negative values indicate lateralization to the left and positive values indicate lateralization to the right. We keep the 0 dB ILD in the repertoire as the sole noise factor. In order to avoid sound artifacts caused by changing the sound pressure level produced by our headphones too quickly, we draw

and apply a new ILD every 25 ms, with a 5% cosine ramp-up and a 5% ramp-down. This results in noise segments that remain constant for 22.5 ms before smoothly changing to a new ILD, unless the new value drawn is the same as the previous in which case the constant period is longer as the figure of the first stimulus we used shows below. We build time-varying stimuli with 8 second duration out of 320 such 25ms segments. Subjects were scored and received visual and auditory feedback of correct and error responses. For the purpose of this feedback, a stimulus was considered as Left or Right according to the mean of the stimulus a subject experienced until reporting her choice, at which time the stimulus was extinguished.

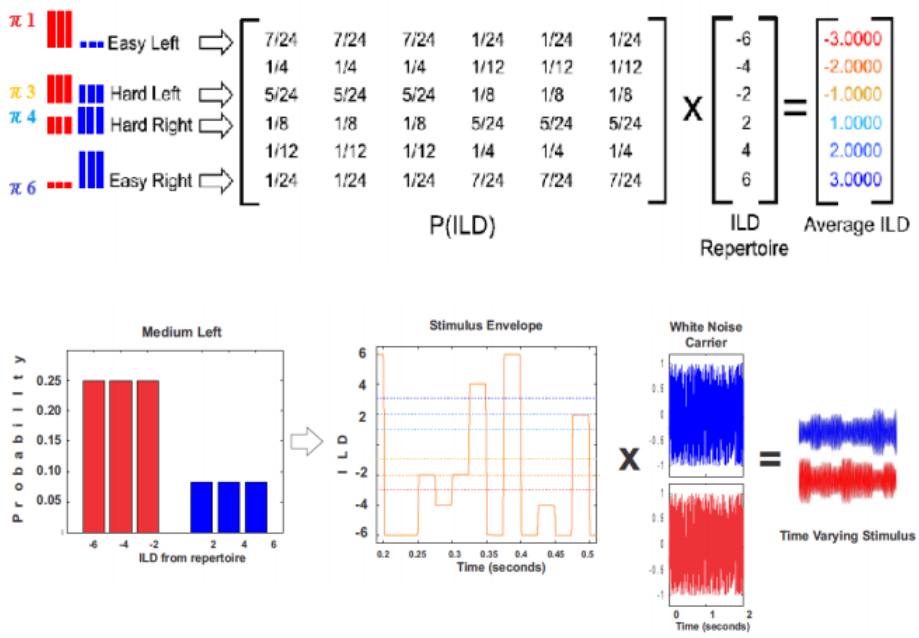


Figure 2.1: We use the interaural level difference (ILD) to create sounds that give the percept of lateralization (left or right). By changing the ILD every 25 milliseconds with controlled probabilities, we create sounds that change their lateralization in time, but are overall louder to one side. The time-varying stimulus envelope is pointwise multiplied by white noise. Top panel: 6 rows of probabilities of sampling from the ILD repertoire determine the mean laterality of the stimulus. Bottom panel: Sampling a sequence of ILDs for medium left difficulty in order to create the stimulus envelope, which is convoluted with white noise in order to produce the stimulus sound waves for each channel.

Finally, to mitigate the effects of stimulus onset and offset on the auditory pathways similar to that of the unwanted auditory artifacts discussed above, we have added 150 ms of non-time-varying cosine ramp-up and ramp-down (see image below). This raises a concern that any auditory deficits to either side could result in biasing the percept of the stimulus towards the opposite side. Subjects for which the point of subjective indifference for sounds with a static ILD is not near 0 dB may decide which side the average lateralization of the entire stimulus favors based on the ramps rather than the time-varying stimulus we are interested in. This is the first motive we have for conducting a pre-testing calibration on each subject to deter-

mine their point of subjective indifference, which we then use to adjust the ramps and stimuli to that particular subject. The improved stimulus used in Experiment 3 (Chapter 4) neglects the need for this calibration.

Figure 2.2: Example of a 3 second sequence of white noise with time-varying sound pressure level (one channel shown), with the ramp up and ramp down periods of 150 milliseconds calibrated to the subject's auditory mid-line at 60 dB SPL.

Although the probabilistic changes featured in our time-varying sound are useful for example generate the linear reconstructions of the stimulus from online motor output, or to conduct reverse correlation analysis (not shown), the stimulus itself sounds very synthetic. The stimulus sometimes transition from one side of the subject's auditory field to the other in only 2.5 ms create two difficulties: 1. Sounds in nature do not behave in this manner, so the perception elicited by this stimulus is not at all one of something moving across the subject's sound space. 2. We want to test whether the motor system will be engaged by this changing stimulus before the subject has arrived at a decision, but the time constant for our stimulus is much shorter (25 ms) than even the blink reflex (100 ms) [Shahani, 1970]. In order to overcome these two problems, we developed a second sound stimulus that preserves the probabilistic nature of the first, but abides to additional criteria concerning the transitions that the sound can make. We employ Markov chains to shape the random walk between ILD states as to give the transitions 1, 2, or 3 degrees of freedom. That is, from the current ILD the sound should only go to an adjacent ILD, to an ILD 2 or 1 step away, or to an ILD 3, 2, or 1 step away. An illustration for transitions with 1 and 2 degrees of freedom can be seen below. We apply the Metropolis-Hastings Chain algorithm to our desired random walks to obtain the transition probabilities that converge to our steady state distributions [Boyd et al., 2004]. Next, we

show an example transition matrix and describe how to use it to construct the random walk stimulus. We begin by choosing the overall lateralization we want to obtain, and the degrees of freedom the transitions between ILDs can have. We next describe the stimulus construction for the medium left condition: We draw one sample from the corresponding distribution, in this case $\pi 2$, and the result will be the current ILD. Say the current ILD is -4, then we would pick the 3rd row from our transition probability matrix and sample once from this distribution. The result becomes the current ILD, and we continue to to draw from the distribution indicated by the transition matrix until 120 samples have been drawn (3 seconds of sound formed by 25ms segments). By chance, some of the constructed sequences will not have converged to the steady state distribution in 120 samples. We generated 10,000 sequences to be able to pick 3 sequences per stimulus condition that are have a maximum $l1$ distance of 10% with respect to their steady state distribution and a maximum difference ± 0.1 dB from the established mean ILD for that condition.

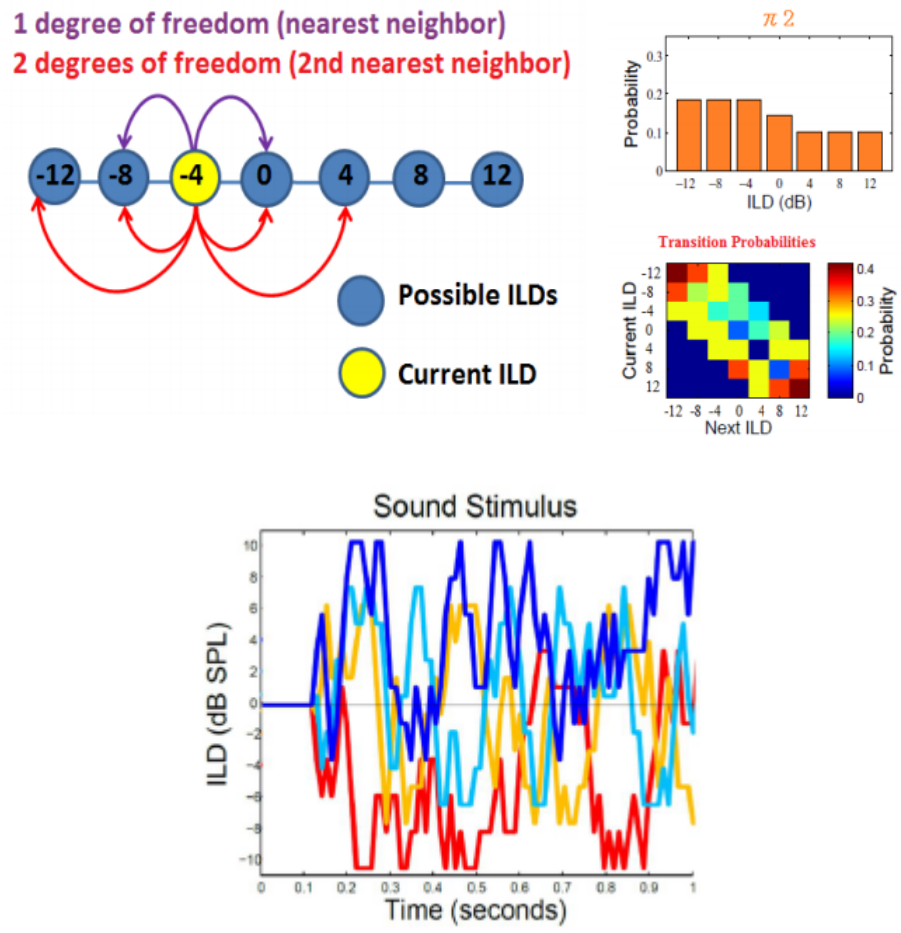


Figure 2.3: Top panels show the construction of sound stimuli that are generated by sampling from a random distribution of interaural level differences biased to negative (left) or positive (right) values. Markov chains are used to compute a transition probability matrix that constrains the step size to jump only to neighboring ILD values in order to create a more continuous stimulus that sounds like the sound source is moving inside the subject's head rather than jumping from side to side. Bottom panel shows sample stimulus sequences color coded by difficulty such that Blue denotes stimuli to the right, Red stimuli to the left, and lighter colors denote harder stimuli.

2.3.3 Online measures of behavior

The haptic joystick utilized in the [Selen et al., 2012] study can be used to monitor online arm movements, but its ability to perturb arm movements comes at a high monetary cost. Since we only desired to measure behavior, we decided to build a simpler set-up such as the one utilized by [Spivey et al., 2005], which could monitor hand trajectories at 60 Hz via computer mouse. We settled for a pc joystick that would similarly allow us to monitor online behavior at 60 hz and display a corresponding cursor through the pc monitor. Eye movements were tracked at 30 fps using a Pupil-Labs eye tracking headset and their open source software previously described [Kassner et al., 2014]. The experiments from Chapter 4 are conducted with a newer version of the tracker capable of 120 fps.

2.3.4 Syncing sounds to other sensors

Great care was taken to ensure that measurements across all the equipment were synced to the sound streams by providing triggers via a NI-DAQ board on the computer running the task. A second computer acquire the NI-DAQ pulses and processed images from video cameras for tracking the arm and the eye of the subject. Bonsai [Lopes et al., 2015] was used for syncing and monitoring all these signals; our novel data is featured in their publication.

Figure 2.4: Top panels shows how triggers produced via NI DAQ or through the sound card itself can be used to provide information about the start of the sound on the computer running the task to a separate computer for online measurements. The middle panel shows how digital triggers can be used to inform the separate computer about the end of a sound. Bottom panel shows a 5ms lag between the execution of an early stop to the stimulus and the effective time that we can change the state of the NI DAQ digital out.

2.4 Results of Experiment 1

2.4.1 Accuracy

For mean ILD values greater than 4 dB to either side, both subjects respond correctly with more than 90 percent accuracy, and mean ILD values greater than 8 dB rarely elicit wrong responses. The psychometric curves confirm that our task, similar as the typical 2-choice task, elicits subjective responses that are more accurate as the stimulus gets easier.

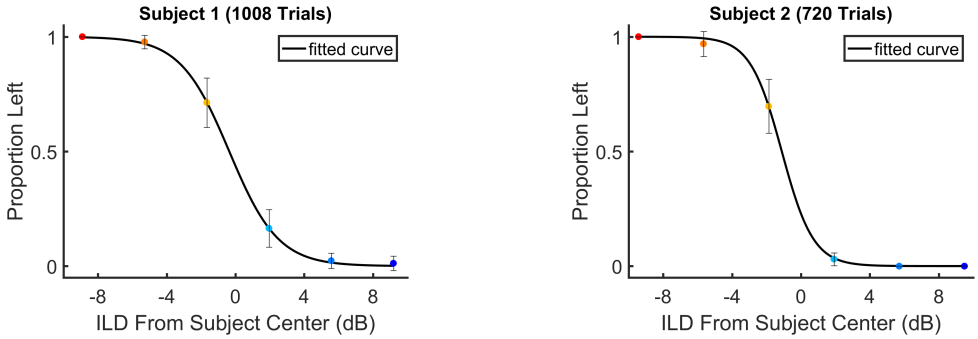


Figure 2.5: Psychometric Curves for subject 1 (left) and subject 2 (right). Post training data is shown for two subjects completing at least 5 sessions of 144 trials. Each session lasted less than 15 minutes. Stimuli were presented in an sound proof booth over calibrated headphones. Colored dots denote the mean proportion of left responses for a particular mean ILD value across all sessions. Error bars show standard deviations computed across sessions.

2.4.2 Reaction Time

Another feature typically observed in these tasks is that easier stimuli, in addition to leading to more accurate responses, also elicit faster response times on average. Analysis of response times in our task reveals this average slow down of responses for harder stimuli, although a linear trend is only significant for subject 1's left oriented stimuli (negative ILD).

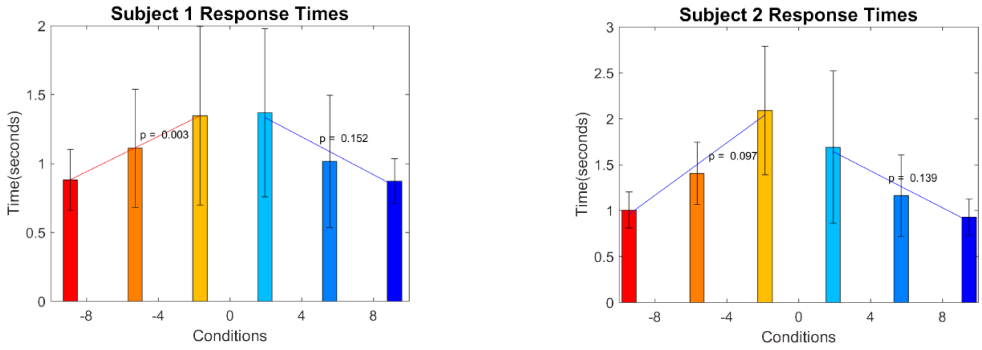


Figure 2.6: Response Times for subject 1 (left) and subject 2 (right). ILD Conditions is defined as the bin centers of the histogram of experienced ILDs with 6 equal bin sizes used to compute the psychometric curves of 2.5. The exact values of ILD conditions can vary between the subjects. Response times appear to decrease as the ILD conditions become easier (darker colors), however, the trend is only significant for subject 1's left sided responses. Error bars show standard deviations.

2.4.3 Choice Trajectory Analysis

Next, we analyze the online behavior of the subjects. Choice trajectories for this experiment were mainly movements direct to the target of choice. However, on a small fraction of the more difficult trials, more complex trajectories going down the middle and changing direction could be observed.

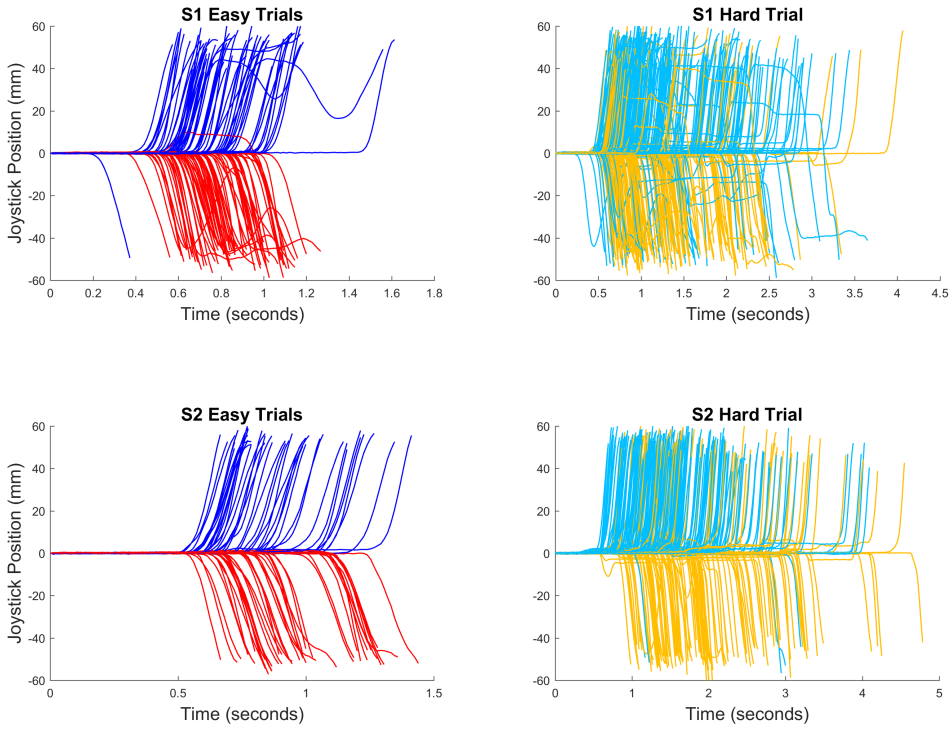


Figure 2.7: Choice trajectories for subject 1 (top) and subject 2 (bottom). Red shifted colors indicate responses to a left biased stimulus while Blue denotes responses to a right biased stimulus, with lighter colors indicating higher difficulty.

One way to summarize this phenomenology is by the length of the choice trajectory for each trial, defined as the cumulative sum of the displacement across the x axis only. When we plot trajectory length as a function of the absolute ILD (collapsing left and right trials), a significant linear decrease of trajectory length with increasing average ILD values across the trial can be observed.

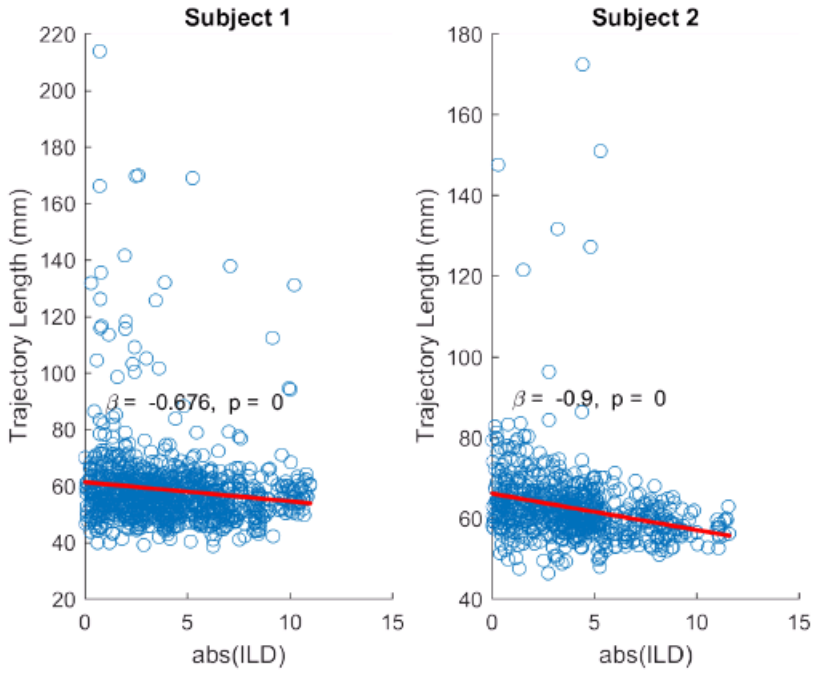


Figure 2.8: Choice trajectory vs stimulus strength (absolute average ILD across the trial) for subject 1 (left) and subject 2 (right). Blue circles are data. Red line shows linear fit with negative weights for both subjects indicating that the Trajectory Length decreases with increasing mean absolute ILD (decreasing difficulty), $p<0.001$.

We investigate this effect further by analyzing correct and incorrect choice trials separately. This analysis reveals that the linear decrease in trajectory length with decreasing trial difficulty is significant for correct trials, but not for incorrect trials for both subjects.

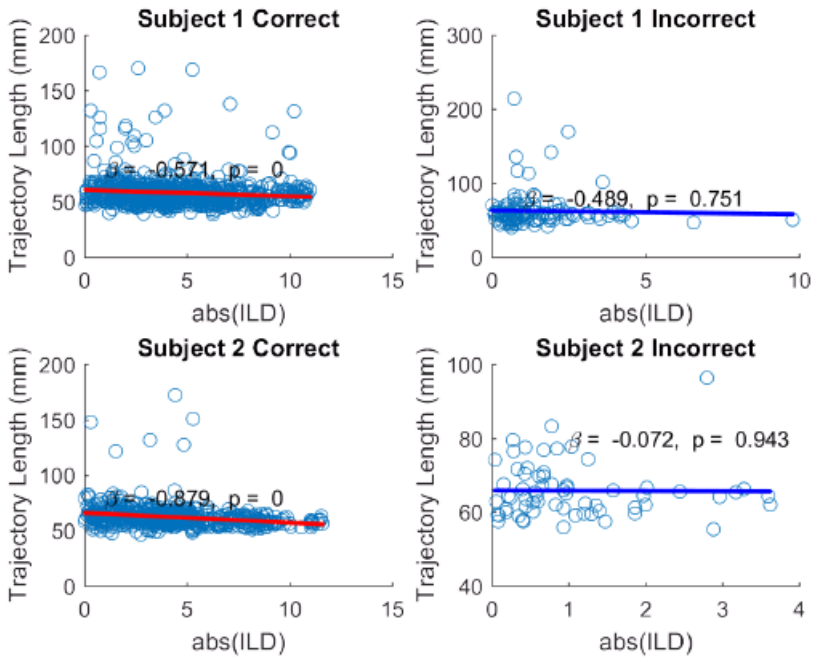


Figure 2.9: Choice trajectory vs stimulus strength (absolute average ILD across the trial) for subject 1 (left) and subject 2 (right). Blue circles are data. For correct trials (left plots), Red line shows linear fit with negative weights for both subjects indicating that the Trajectory Length decreases with increasing mean absolute ILD (decreasing difficulty), p<0.001. For incorrect trials (right plots), Blue line shows linear fit with negative weights for both subjects indicating that the Trajectory Length decreases with increasing mean absolute ILD (decreasing difficulty), Not Significant.

2.5 Discussion of Experiment 1

This first experiment showed us that it is possible to extend the standard two-choice task to obtain behavior during the deliberation period by incorporating a spatial component. Stable psychometric curves across several sessions and the pattern of increased reaction times with increasing trial difficulty are important, since they tell us that final choice behavior in this

task resembles that of the standard task. This is meaningful since it enables proper statistical analysis of the data, and because any insights we draw are best interpreted in light of classical and modern psychophysics research. Although previous studies had already incorporated such a spatial component to the binary choice task [Spivey et al., 2005, Selen et al., 2012], and stochastic time-varying stimuli [Nienborg and Cumming, 2007], to our knowledge this is the first time that a stochastic time-varying stimulus was presented while subject behavior was densely measured.

To the bare eye, the most salient feature of our data is perhaps the occurrence of hesitant choice trajectories elicited during some hard trials. This phenomenology finds a natural summary statistic in the length of the choice trajectory. The fact that trajectory length was only significantly correlated to the overall mean of the stimulus for correct trials might be a hint that our online measures of behavior hold information about the subject's confidence that they are making the right choice. Different patterns of behavior for correct vs incorrect trials have been linked to confidence by previous studies [Zariwala et al., 2013], but this has typically been studied from reaction times, which decrease as stimulus difficulty decreases for correct trials whereas incorrect trials have the opposite profile. However, a separate analysis of reaction times for correct vs incorrect trials did not reveal this effect for our task (this negative result is not shown).

Although these initial results were exciting to us, we found that the number of complex trajectories per subject were too few for more meaningful analysis, motivating us to tweak the task such that more complex trajectories could be observed (Figure 2.17). Our follow up experiment is detailed next.

2.6 Motivation for Experiment 2

While we managed to create a decision making task in the auditory domain in which behavior could be tracked beyond accuracy and reaction time during the discrimination period, the number of hesitant decision trajectories were too few for meaningful analysis. We wanted to observe more trajectories of this type, since we thought they might reflect online information about the state of the subject's decision. We suspected that a sense of urgency to move would be needed to match a subject's movements with the faster time-scale of the stimulus and the subject's moment-to-moment be-

liefs about it. We thus decided to add the constraint that subjects should initiate their choice trajectories within 500 milliseconds of the start of the stimulus, else the trial was aborted. This modification resulted in a much higher number of hesitant trajectories per session for Experiment 2 relative to Experiment 1. The new data allowed us to statistically explore the relationship between the time varying stimulus and the online movements. We found that linear combinations of eye and hand signals allow the reconstruction of the temporal structure of the stimulus. Understanding how this coupling arises required a more theoretical approach, the topic of Chapter 3, which in itself led to further change the task described in Chapter 4.

2.7 Methods of Experiment 2

2.7.1 Subjects

Subjects received 3 training sessions prior to entering the experiment. A total of 5 naive subjects were considered for this experiment, but 1 subject was excluded because he failed to achieve a 70 percent correct performance by the 3rd session, all other 4 subjects went onto the experiment. All subjects performed 1080 trials across 12 sessions (3 sessions per day with a small break in between). Psychometric curves constructed from these 12 sessions are shown for all subject in the results section.

2.7.2 Task

Experiment 2 had two modifications relative to Experiment 1. First, in order to increase the dynamic range or the arena, subjects now had to move their arm forward (y-direction of arena). We thought requiring reach-like movements to complete the task could lead to higher chances of engaging the motor system during the task. Second, in order to impose a sense of urgency, subjects were instructed to begin moving within 500 milliseconds of the start of the sound. Failure to move within 500 milliseconds resulted in an annoying buzzing sound and the screen markers flashing red. Subjects were free to take as long as 8 seconds (the maximum stimulus duration) to report their final choice.

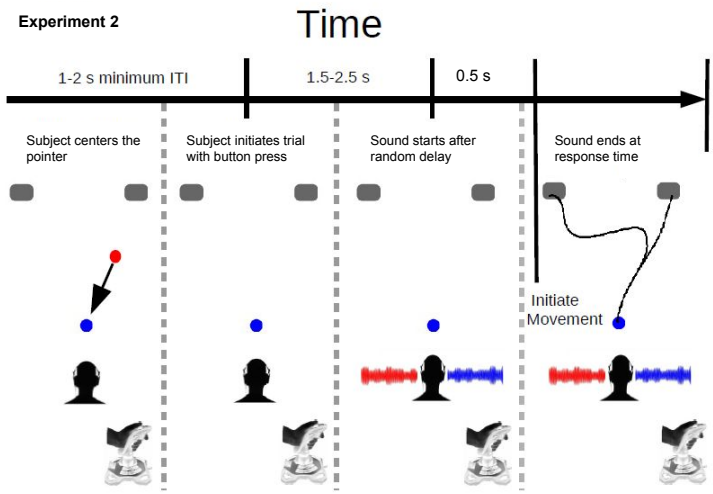


Figure 2.10: Schematic of temporal sequence of events in the task.

2.7.3 Stimulus

The stimulus is as the colored stimulus described in Experiment 1 2.3, but with one small change. The steady-state distribution from which sounds were sampled was changed to a uniform distribution of ILDs from -6dB to +6dB, which was shifted by 0.5, 2, or 4 dB to negative values to induce a left bias and to positive values to induce right bias. This modification resulted in qualitatively similar sounds as before, but the sampled ILD distribution converges to the theoretical mean of the stimulus slightly faster. Subjects were scored and received feedback according to the mean of the stimulus they experienced until reporting their choice, at which time the stimulus was extinguished. An example of an easy left trial with this slightly modified stimulus is shown. The figure shows a distribution shifted by -3 dB, however in the experiment the distribution was shifted by -4dB to make the stimulus even easier.

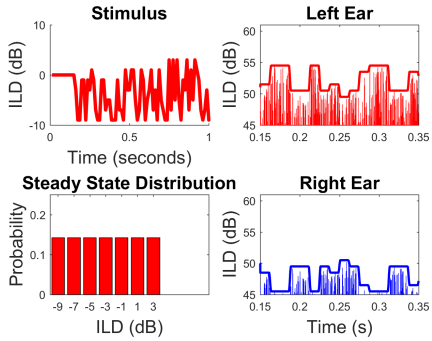


Figure 2.11: Top left: Sample easy left stimulus. Bottom Left: Uniform steady state distribution the stimulus is sampled from. Top right: Left channel white noise stimulus. Bottom right: Right channel white noise stimulus.

2.8 Results of Experiment 2

2.8.1 Accuracy

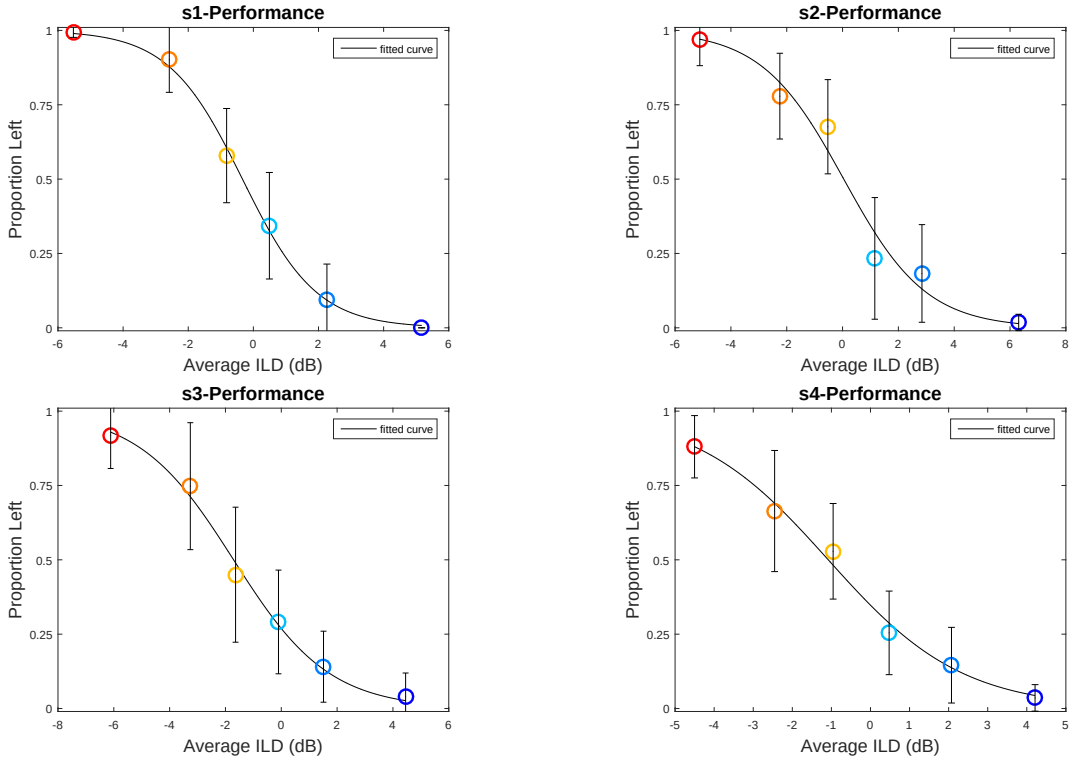


Figure 2.12: Psychometric curve for all subjects computed from 1080 trials post-training. Colored dots denote the mean proportion of left responses for a particular mean ILD value across all sessions. Error bars show standard deviations computed across 12 sessions.

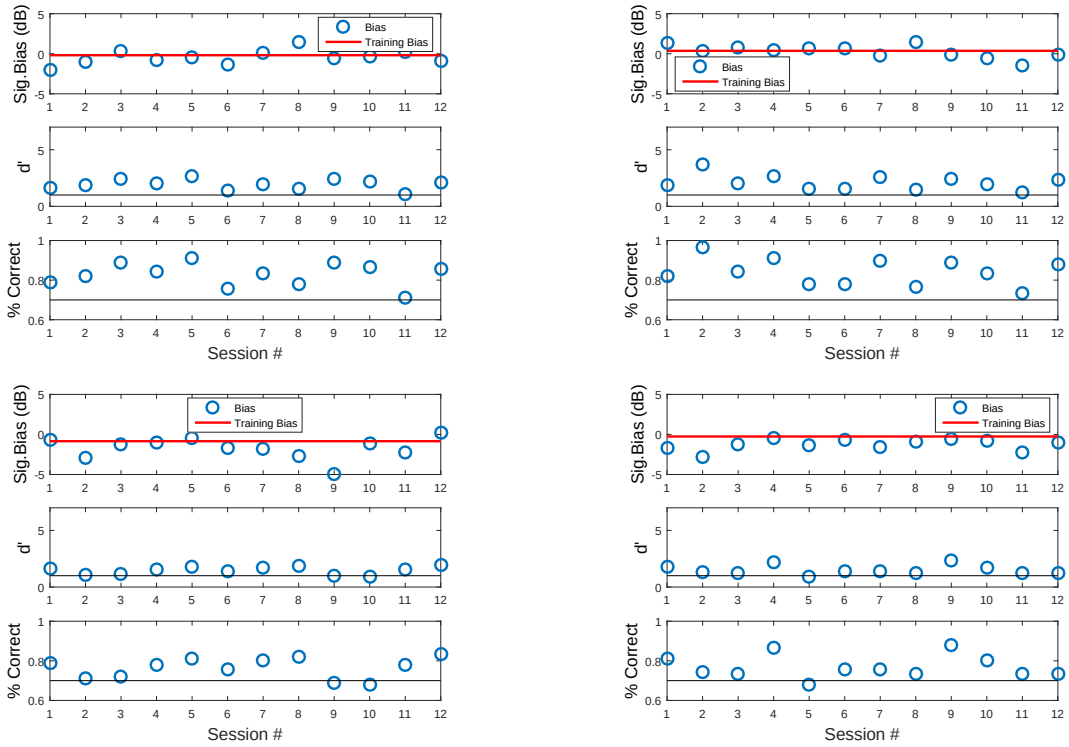


Figure 2.13: Four panels with 3 rows each show the performance of each subject across sessions. Psychometric curve bias (top), d' (middle) and accuracy (bottom) shows performance remained relatively stable across sessions and days. For the d' plots, the horizontal line marks d'=1. For the Percent Correct plots, the horizontal line marks 70 percent accuracy. Note that subjects 1 and 2 never fall below these performance thresholds, while subjects 3 and 4 were below only 2 and 1 sessions respectively.

2.8.2 Reaction Time

Forcing subjects to start moving within the first half second caused subjects to have a sense of urgency that was not present in the first experiment, reflected in longer Reaction Times (RT), the time when the trajectory ends. When we analyze RT as a function of the actual experienced stimulus, a significant decrease in RT with increasing bias of the stimulus can be observed for all subjects. Red (blue) traces show leftward (rightward) stimuli, with lighter colors showing easier trials.

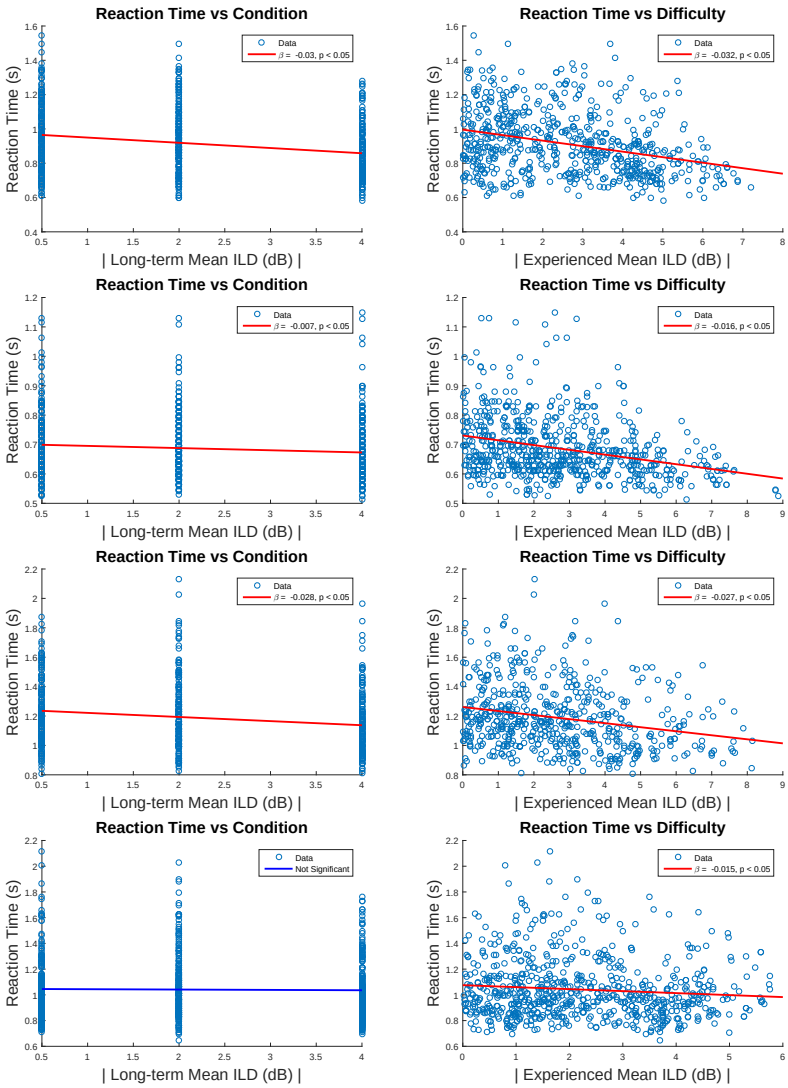


Figure 2.14: Reaction time as a function of the long-term mean of the stimulus (left) and the experienced mean of the stimulus (right) for each subject per row. Blue circles show data, red lines linear fits with significant absolute mean ILD coefficient, and blue lines linear fits that don't have a significant absolute mean ILD coefficient. The plots on the right column show a significant linear decrease in reaction time with increasing absolute experienced mean trial ILD (decreasing difficulty) for all subjects. The decrease in reaction time with decreasing absolute long-term mean of the stimulus is also significant for 3 of the 4 subjects.

2.8.3 Trajectory Analysis

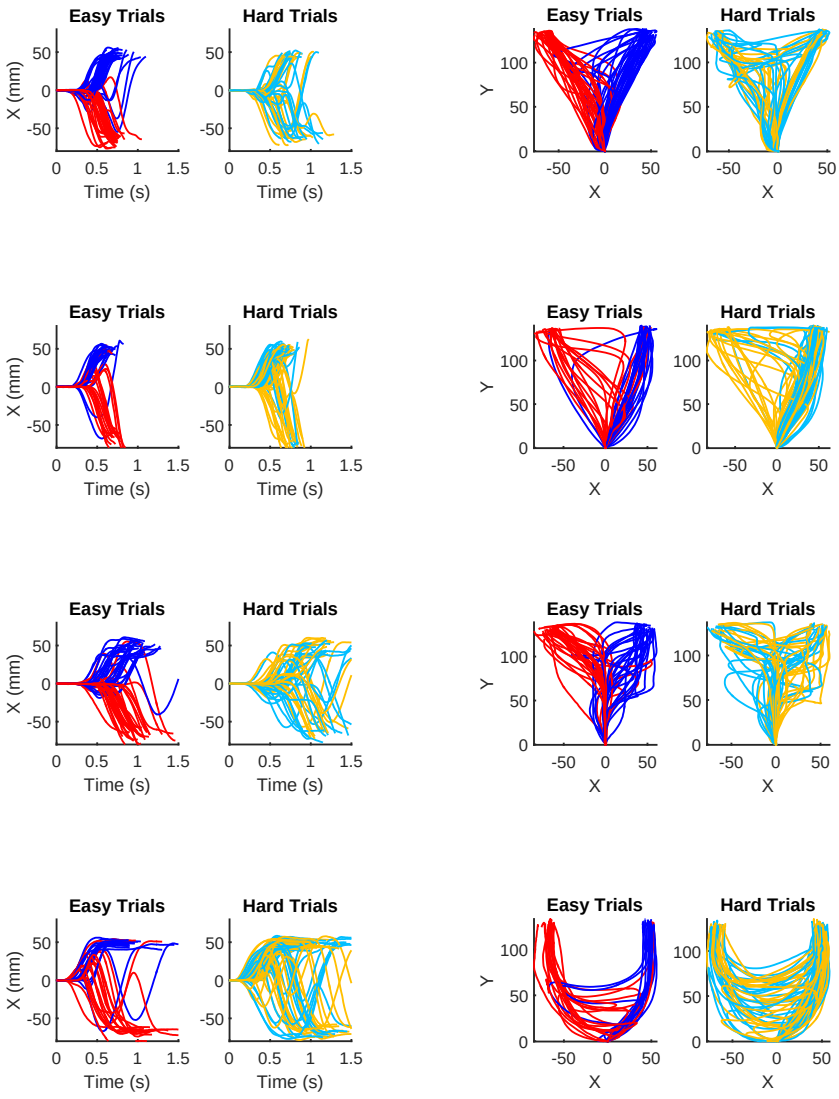


Figure 2.15: Experiment 2 trajectories from each subject's last 150 trials. Each row is one subject. Left 2 columns show the x-axis in time. Right 2 columns show the trajectories in xy.

In order to compare the number of hesitant trajectories between the experiments, it was necessary to split the hesitant trajectories from the more ballistic or direct to target trajectories for each subject. For experiment 1 this was easily accomplished by setting a threshold on trajectory length. However, the extension of the task to xy coordinates and the 500 millisecond limit to begin a trajectory resulted in subjects developing unique strategies that make it hard to classify stereotyped vs. non-stereotyped behavior. For this reason, we used k-means clustering to divide trajectories into 2 categories on the basis of kinematic parameters of the each trajectory. Then the cluster with lower (higher) standard deviation in trial trajectory position is labeled as the stereotyped (hesitant) trial type.

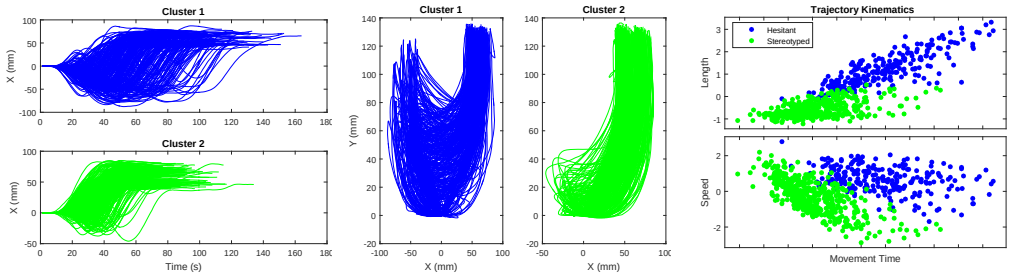


Figure 2.16: Clustering of experiment 2 subject 4 data into stereotyped vs hesitant trajectories using k-means with 2 clusters on the trajectory parameters (movement start time, trajectory length, trajectory speed, and acceleration). Note all left ending trajectories were flipped to end on the right prior to clustering for ease of visualization. Left panel shows the trajectories in time, center panel shows them in the xy coordinates of the arena, and the right panel shows the resulting clusters in Length (top) and Speed (bottom) vs Movement Time

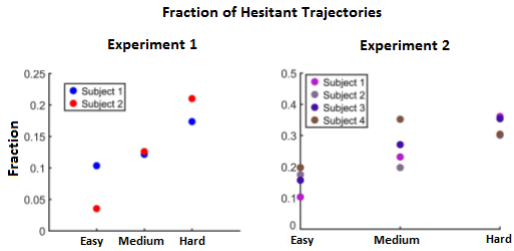


Figure 2.17: We observed more hesitant trajectories for experiment 2 (right panel) than for experiment 1 (left panel). Color code of legend identifies each subject.

As before, we plot the trajectory length as a function of the absolute mean ILD value for each trial. A significant increase in trajectory length with decreasing absolute mean ILD values (increasing difficulty) can be observed for all subjects.

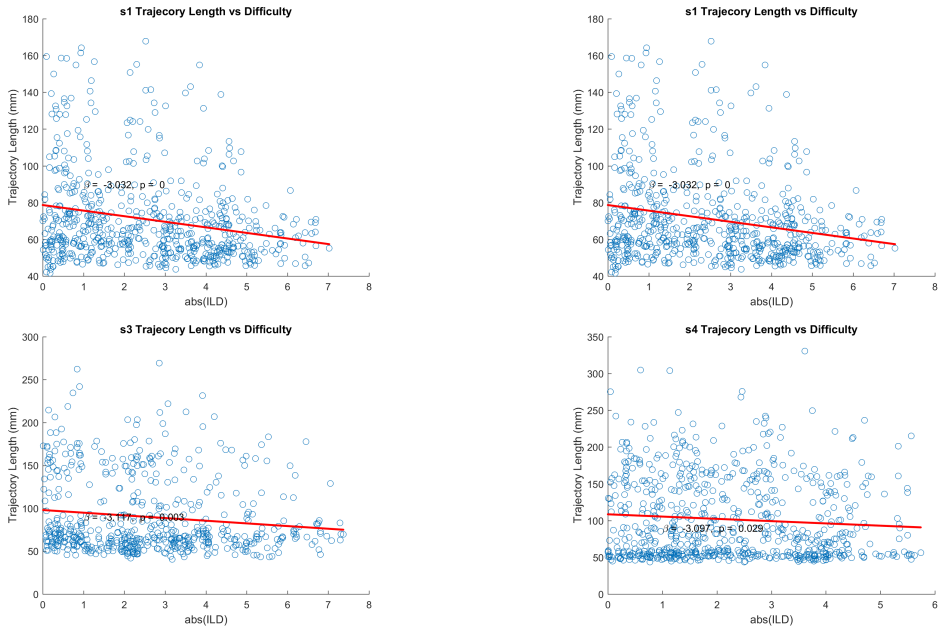
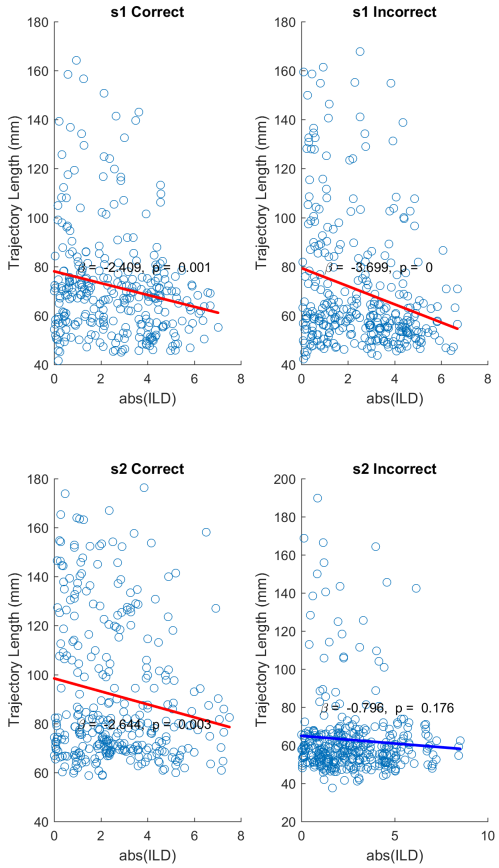


Figure 2.18: For all subjects we observed a significant increase in trial trajectory length with decreasing absolute mean ILD values (increasing difficulty). Blue circles show data, red lines show linear fits with significant coefficient for the independent variable.

We once again conduct the same analysis separately for correct and incorrect trials, but we do not observe any consistent difference in trajectory length between correct and incorrect trials across subjects. While in experiment 1 the increase in trajectory length with increasing difficulty was significant only for the correct trials, results for experiment 2 are inconclusive as only subject 2 shows the same pattern, subject 4 has a significant effect only for incorrect trials, and subject 1 and 3 show a significant effect for both correct and incorrect trials.



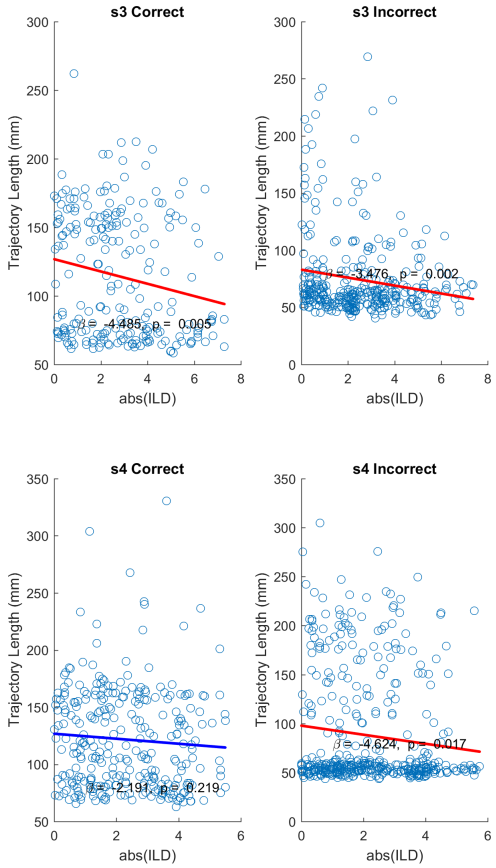


Figure 2.19: Trajectory length analysis for correct and incorrect trials separately. The increase in trajectory length with increasing trial difficulty is significant for subject 1 and subject 3 for both correct and incorrect trials, while subject 2 shows it only for correct trials and subject 4 has a significant effect only for incorrect trials. The results of this analysis are inconclusive regarding confidence effects. Blue circles show data, red (blue) lines show linear fits with significant (not significant) coefficient for the independent variable.

2.8.4 Stimulus Reconstructions

The goal of the following analysis is to evaluate whether information about the stimulus is present in the online eye and hand trajectories, by reconstructing the stimulus from linear combinations of these online readouts of

behavior. Regularized (ridge) linear regression of the moment-to-moment stimulus onto the eye and hand data was performed, using for each stimulus segment to be predicted a window of 500 milliseconds of eye and hand data, which moves forward with the evolving stimulus. Ridge regularization was applied due to the co-linearity problem arising from correlations in the eye and hand data. Ridge regression minimizes squared error while regularizing the norm of the weights by parameter λ:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i} (\beta^T x_i - y_i)^2 + \lambda \beta^2. \tag{2.1}$$

Training and testing was performed using the 'hold-out' method, using 60 percent of the clean data to train the model and obtain a prediction coefficient kernel for each subject. The linear kernels were then used to reconstruct the stimulus from the remaining 40 percent of eye and hand data never seen by the model, and the correlation between the resulting test set reconstructions and the actual stimuli was computed. The regularization coefficient was visually chosen to maximize the mean correlation between reconstructed and actual stimuli for each subject's test set. This method only produced reasonable kernels and predictions restricting it to the difficult trials (experienced trial mean ILD < 0.1 dB).

A schematic of how each linear decoder (one per time-step) is trained with a window of eye and hand data is shown below.

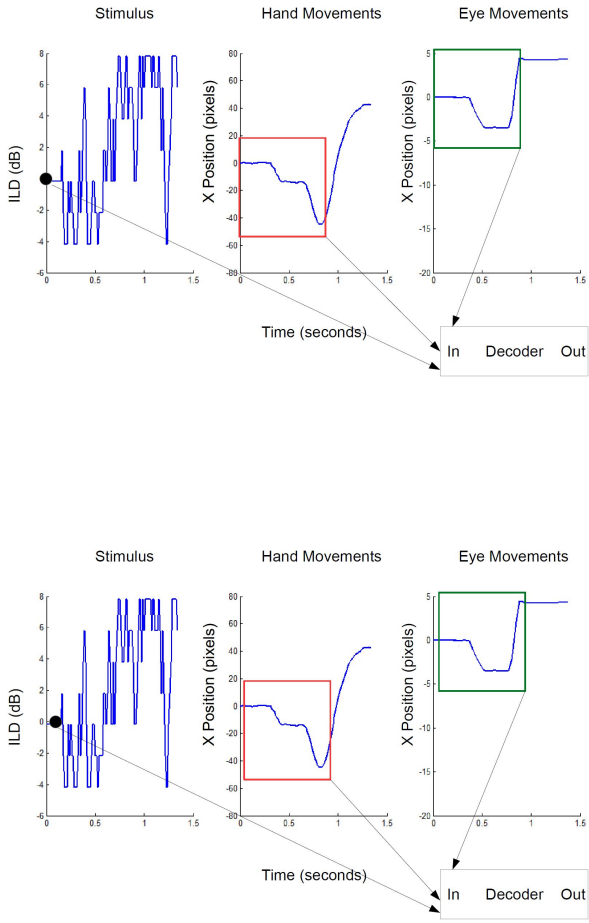


Figure 2.20: Schematic of training a decoder to predict the moment-to-moment stimulus from a window of eye and hand data. For each stimulus segment to be predicted a window of 500 milliseconds of eye and hand data, which moves forward with the evolving stimulus. Top panel shows the target variable (stimulus), and the corresponding predictor windows (Hand and Eye) for time-step 1. Bottom panel shows the same variables for time-step 2.

The resulting kernels for all subjects are show in (Figure 2.21). In each panel the top half of the plot, each slice of time shows the linear weight

of each point in a window of the next 500 milliseconds of hand data - at the bottom half the corresponding linear weight of windows of eye data - for predicting the instantaneous ILD value at each time following the onset of each stimulus sample. In subjects 1, 3, and 4, prior to the 700 ms time following stimulus onset, patches of negative (blue) can be observed below patches of positive (red) weights can be seen, especially in the hand kernels (top). This means that hand movements are not equally informative about the stimulus ILD at different times, and is evidence that subjects are not integrating information equally across time in order to guide their hand movements. Furthermore, the temporal resolution of the stimulus allows us to see that for the hard trials considered in the analysis, hand position is on average negatively correlated with ILD value for 100-300 milliseconds and then correlates positively for the next 100-300 ms. This suggests that subjects are listening for sweeps in the stimulus from one side to the other prior to moving their hand in the direction of the sweep. Alternatively, this pattern might be a result of the changes of direction, presumably following sequences of changes of mind we observe in the response trajectories. In either case, this is empirical evidence that subjects are tracking the stimulus with their hands and eyes. The coefficients for eye data tend to be weaker, and are more difficult to interpret, but similar but smaller patches as we observed for the hand kernels appear to be present in the eye kernels too. In (Figure 2.22) we show a few samples of the reconstructed stimuli, including both good and poor reconstructions. We then show histograms of the correlation between the actual stimuli and the linear reconstructions in hold-out test data (Figure 2.23). We think the fact that the mean correlation of these reconstructions to the true stimulus is greater than zero ($p<0.001$) is meaningful, because it shows that even this very limited model can, to some degree, accurately encode and decode the stimulus. The only subject for which the analysis is not significant is subject 2, which is also the subject for which the stimulus kernel becomes uninformative earliest (by 500 ms vs 700 ms for the other subjects). Qualitatively, subject 2 appears to have the most direct to target trajectories and rarely shows more than one change of mind (Figure 2.15). This qualitative observation is consistent with the conclusions we reached that this subject is not tracking the stimulus, per the more quantitative kernel and reconstruction analysis.

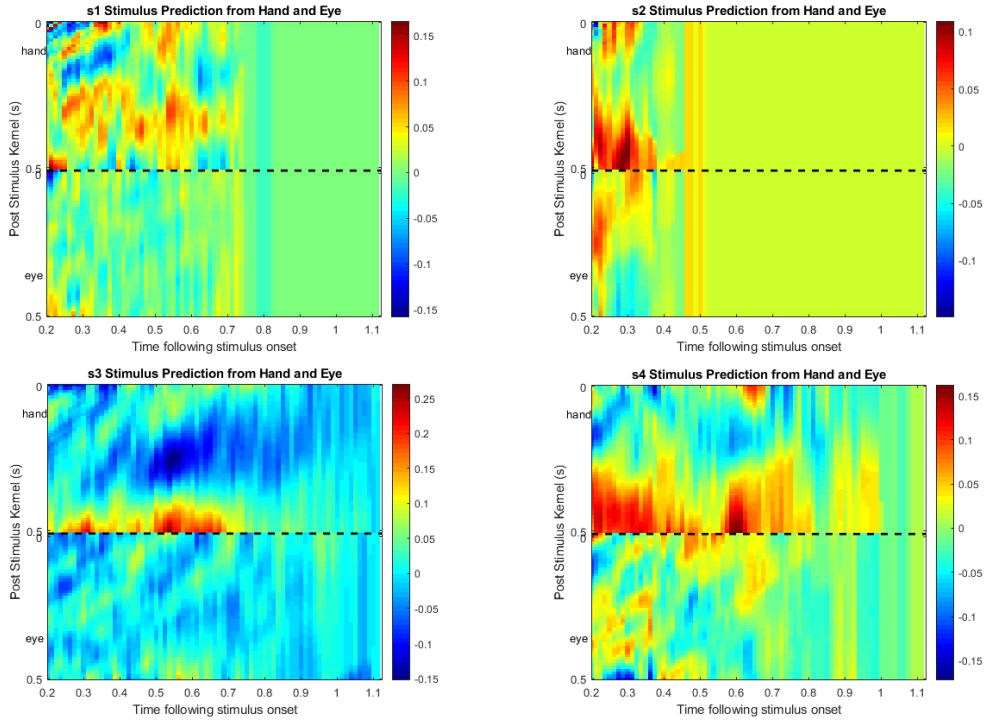


Figure 2.21: Regularized (ridge) linear regression of the moment-to-moment stimulus onto the eye and hand data for all subjects. Regularized (ridge) linear regression of the moment-to-moment stimulus onto the eye and hand data was performed, using for each stimulus segment to be predicted a window of 500 milliseconds of eye and hand data, which moves forward with the evolving stimulus. Post-stimulus kernel refers to the prediction coefficient (β) of each sample in each window of hand and eye data. For each plot, the top panel shows the prediction coefficients kernel fit on 60 percent of the data (training set) for the hand data and the bottom panel eye data. The hand and eye data were z-scored prior to fitting the model.

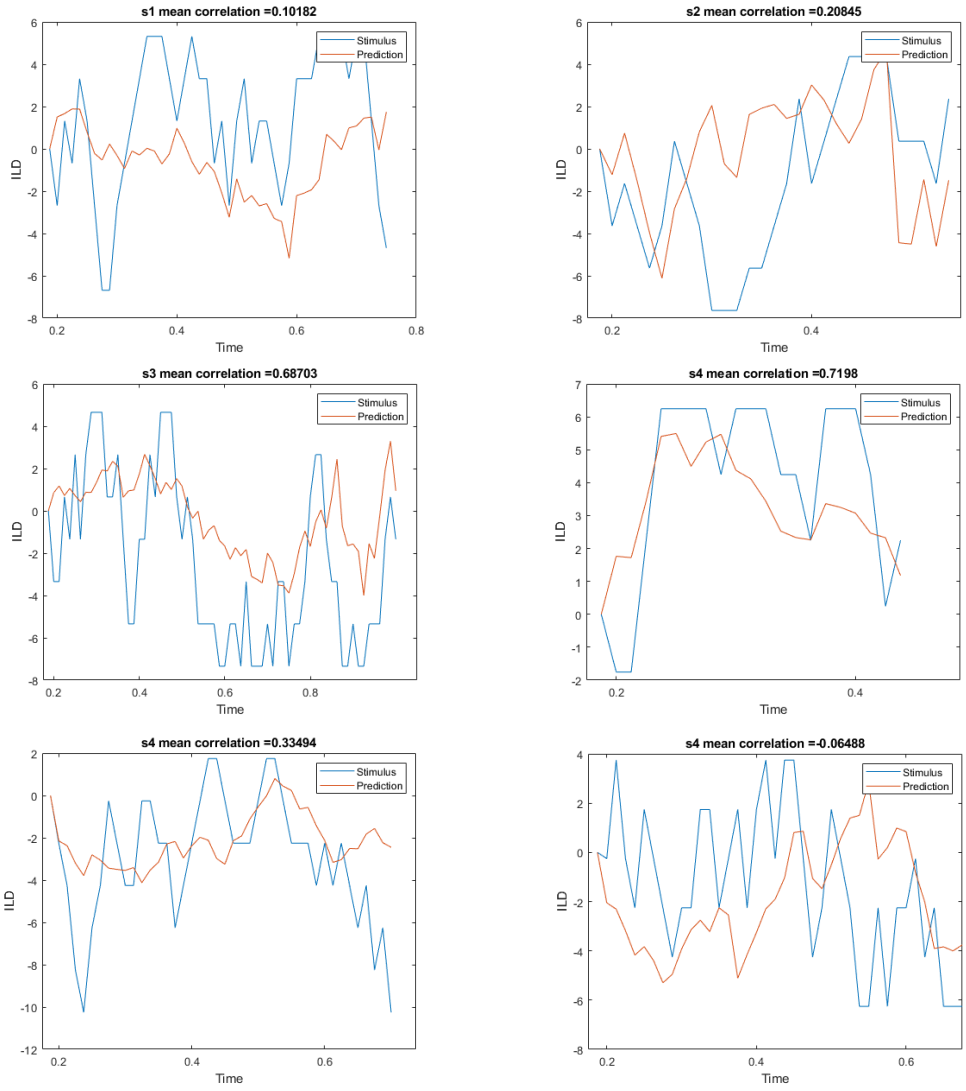


Figure 2.22: Reconstructed stimuli vs. actual stimuli sampled from the test set, not used during training. The title for each plot shows the correlation between the two time-series. Legend: Blue line shows the stimulus, and red line the reconstruction from arm and eye data.

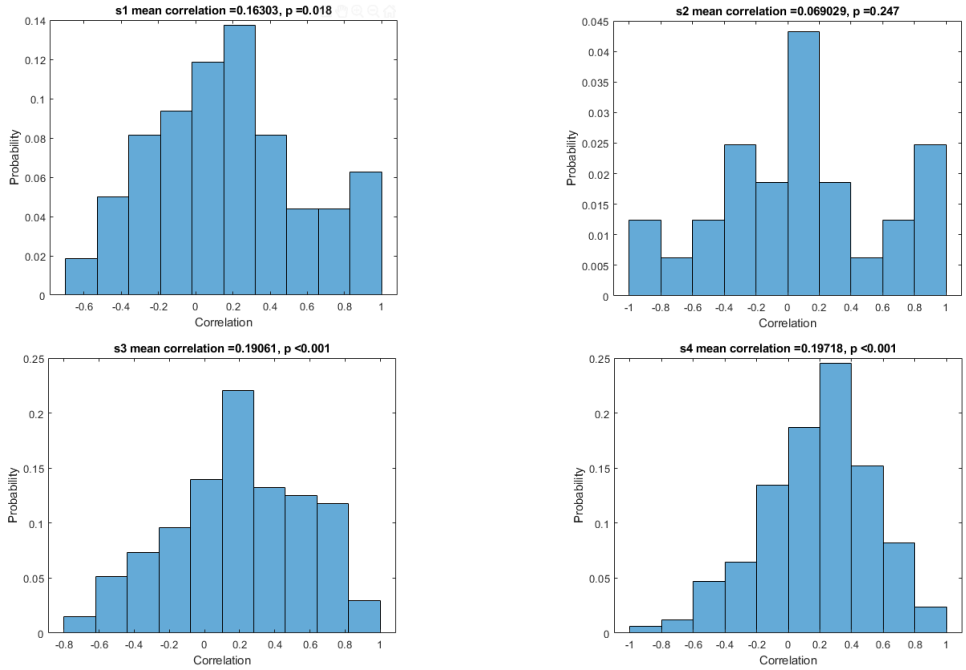


Figure 2.23: Histograms of the correlation between actual and reconstructed stimuli for all subjects for the test-set. A p-value was calculated for each mean correlation coefficient by shuffling the varying part of the actual stimuli in order to create a distribution of surrogate mean correlations to the reconstructions. Mean correlation for all subjects, except subject 2, was greater than zero and significant. Subject 2 (top right) has the lowest mean correlation, which is also not significant. This is to be expected from her rather ballistic trajectories, faster reaction times, and prediction coefficients kernel which cuts off after 400 milliseconds into the stimulus.

2.9 Discussion

Previous studies managed to obtain online readouts of behavior during the execution of a 2AFC task, for example using a computer mouse [Spivey et al., 2005] or a haptic joystick like controller [Resulaj et al., 2009]. In both of these studies, hard trials elicited complex decision trajectories. One study described these complex decision trajectories as 'in-between' [Spivey et al., 2005], and the other as changes of mind [Resulaj et al., 2009]. The authors of

changes of mind in decision making could be predicted using the bounded accumulation of evidence diffusion model. We were inspired by these studies, and wanted to extend the methodology to investigate further the relationship of online movements during the task and the stimulus. The hard stimuli used for the discrimination task in [Spivey et al., 2005] were phonological competitors such as 'candle' and 'candy' (vs 'dolphin' for an easier discrimination), which are not suitable for studying the evolution in time of the mental processes we are interested in. Statistically constructed stimuli such as the one used in [Nienborg and Cumming, 2007, Resulaj et al., 2009] is more suitable for our purposes, but in those studies they extinguish the stimulus at movement initiation. Later work by Shadlen's group established a correlation between the online stimulus and motor signals recorded prior to the movement initiation [Selen et al., 2012]. We managed to obtain online motor responses during the accumulation period in our first experiment, but this behavior was only observed in a fraction of the more difficult trials. Our second experiment had the explicit goal of increasing the number of hesitant trajectories performed by the subjects in our novel binary choice task. In experiment 1, we thought that subjects were waiting too long to begin moving, allowing for their decision to evolve to a fully committed choice by the time most actions were initiated. In experiment 2, imposing a cut-off time of 500 milliseconds to begin movements was enough to double the number of hesitant trajectories observed per session. However, this constrain also resulted in shorter reaction times and thus less time for the subjects to experience and report on the stimulus. Our novel task allowed us to quantitatively assess the degree to which online behavior encodes information about an ongoing binary discrimination. Like Spivey [Spivey et al., 2005], we believe our online readouts to be empirical evidence for continuous uptake of sensory input and dynamic competition between simultaneously active representations (of speech lexicons in their study, of left and right ILDs or responses in our case). However, the controlled time-varying nature of our stimulus allowed us go one step further in our analysis and quantitatively estimate the amount of information about the stimulus that is encoded in the online behavioral readouts. We found that linear combinations of eye and hand data can be used to reconstruct the online stimulus. The predictive power of our model, demonstrated through cross-validated reconstructions in (Figure 2.23), serves as validation while its simplicity facilitates the interpretation of precise stimulus driven events that can be seen in the kernels. For example,

we interpret the sharp drop in the prediction coefficient kernels around 700 ms for subjects 1, 3, and 4 (Figure 2.21) as a switch in subject behavior from seemingly tracking the stimulus to ignoring it as they commit to a choice. The patches of negative (blue) transitioning sharply to positive (red) prediction weights in the hand kernels show that subjects are utilizing stimulus information to guide their hand movements differently across time, however, it is hard to distinguish whether these patterns are due to some heuristic being applied by subject to process the stimulus, such as detection of sweeps from one side to the other, or are instead due to changes of mind. We initially thought that the transition from seemingly tracking the stimulus to ignoring it (Figure 2.21) was a result of a non-linearity in the processing of the stimulus, reflecting the categorical nature of committing to a choice. However, the data did not allow us to exclude the hypothesis that the non-linearity could arise from the motor component to the task (e.g. the need to control an arm). As we tried to formalize our description of subject behavior in this task, the question 'Why are subjects behaving in this manner in our task' turned into 'How should subjects behave in this task?' In order to answer this question, we built normative models for our task by expanding upon a normative model of the standard two choice task [Drugowitsch et al., 2012]. The original framework models an optimal agent's urgency to act by subtracting a cost for time passing from the agent's notion of long-term rewards in the task. Our framework generalizes the model to embodied settings requiring an agent to move an arm or navigate an environment, and is the subject of the next chapter. Ultimately, we learned that time urgency was a necessary component in order to observe the uncommitted behavior we were interested in studying, and it became clear that we needed to be able to control the subjects' urgency in the task. This led to the implementation of a point system for the task of Chapter 4 that tries to link the explicit rewards and cost for time passing of the models of chapter 3 to monetary compensation for the subjects of the experiments of Chapter 4. In experiment 3, the subjects receive visual feedback regarding expected rewards and points lost due to the passage of time during the execution of the task, and summary of the accumulated totals between trials. The difficulty of analyzing subject behavior in terms of the experienced stimulus (Figure 2.21) forced us to rethink the stimulus as well. The fast features of the stimulus are not clearly translating to behavior, with longer sequences of stimuli (patches of blue or red) to one single side appear to be what guides subject behavior.

The regions of positive and negative weighs patterned after each other might suggest that subjects are also tuning to sweeps from one side to the other in the stimulus. We wanted to have a simpler time-varying stimulus for the next round of experiments. We were inspired by an auditory task that used sequences of binary (left or right) clicks in order to study decision making in rats and humans [Brunton et al., 2013]. In the [Brunton et al., 2013] study, auditory clicks are generated from independent Poisson processes, and the Poisson click rates are constant for the duration of each trial. The difference in click rates sets task difficulty. We instead use a constant click-rate (4 Hz) and modulate a single probability of a click occurring on the left or the right, and that sing probability controls task difficulty.

2.10 Acknowledgments

Experiment 1 data was collected and analyzed by me with the guidance of Jose L. Pardo-Vazquez and Alfonso Renart. For experiment 2, we thank Gautam Agarwal for his contribution in providing an innovative approach to analyzing our data and for discussions, and Jose L. Pardo-Vazquez and Zach Mainen for discussions.

Bibliography

[Boyd et al., 2004] Boyd, S., Diaconis, P., and Xiao, L. (2004). Fastest mixing Markov chain on a graph. *SIAM Review*. (Cited on page 24.)

[Brunton et al., 2013] Brunton, B. W., Botvinick, M. M., and Brody, C. D. (2013). Rats and humans can optimally accumulate evidence for decision-making. *Science (New York, N.Y.)*, 340(6128):95–98. (Cited on pages 54 and 145.)

[Drugowitsch et al., 2012] Drugowitsch, J., Moreno-Bote, R. N., Churchland, A. K., Shadlen, M. N., and Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. *Journal of Neuroscience*. (Cited on pages 6, 7, 9, 53, 58, 59, 60, 62, 65, 66, 74, 76, 78, 97, 98, 106, 153, and 154.)

[Kassner et al., 2014] Kassner, M., Patera, W., and Bulling, A. (2014). Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction. In *UbiComp 2014 - Adjunct Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. (Cited on page 27.)

[Lopes et al., 2015] Lopes, G., Bonacchi, N., Frazão, J., Neto, J. P., Atallah, B. V., Soares, S., Moreira, L., Matias, S., Itskov, P. M., Correia, P. A., Medina, R. E., Calcaterra, L., Dreosti, E., Paton, J. J., and Kampff, A. R. (2015). Bonsai: An event-based framework for processing and controlling data streams. *Frontiers in Neuroinformatics*. (Cited on page 27.)

[Nienborg and Cumming, 2007] Nienborg, H. and Cumming, B. G. (2007). Psychophysically measured task strategy for disparity discrimination is reflected in V2 neurons. *Nature Neuroscience*. (Cited on pages 34, 52, and 145.)

[Pelli and Farell, 1996] Pelli, D. G. and Farell, B. (1996). Chapter 29 - Psychophysical Methods. *Handbook of Optics*. (Cited on page 19.)

[Rajkowski et al., 1994] Rajkowski, J., Kubiak, P., and Aston-Jones, G. (1994). Locus coeruleus activity in monkey: Phasic and tonic changes are associated with altered vigilance. *Brain Research Bulletin*. (Cited on page 18.)

[Resulaj et al., 2009] Resulaj, A., Kiani, R., Wolpert, D. M., and Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature*. (Cited on pages 8, 9, 51, 52, 145, 151, and 152.)

[Selen et al., 2012] Selen, L. P., Shadlen, M. N., and Wolpert, D. M. (2012). Deliberation in the motor system: Reflex gains track evolving evidence leading to a decision. *Journal of Neuroscience*. (Cited on pages 9, 20, 27, 34, 52, 145, and 151.)

[Shahani, 1970] Shahani, B. (1970). The human blink reflex. *Journal of neurology, neurosurgery, and psychiatry*. (Cited on page 24.)

[Spivey et al., 2005] Spivey, M. J., Grosjean, M., and Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences of the United States of America*. (Cited on pages 8, 20, 27, 34, 51, 52, 145, 151, and 152.)

[Zariwala et al., 2013] Zariwala, H. A., Kepecs, A., Uchida, N., Hirokawa, J., and Mainen, Z. F. (2013). The limits of deliberation in a perceptual decision task. *Neuron*. (Cited on page 34.)

Chapter 3

Optimal decision making model for temporally extended actions

HIGHLIGHTS

- In this chapter we describe a normative reinforcement learning agent that shows hesitant decision trajectories for difficult stimuli for a range of costs for time passing and for moving.
- The model's average trajectory length increases significantly with increasing stimulus difficulty (decreasing mean ILD).
- Increasing the cost of time (decreasing time limit) causes the agent's (soft) action bounds to narrow, such that less extreme beliefs are required for the agent to move rather than wait.
- Increasing the width of the arena shifts the agent's (hard) decision bounds to earlier times, thus committing the agent to a choice earlier.

Contents

3.1 Abstract **58**
3.2 Introduction **58**
3.3 Methods **60**
3.4 Results **74**
3.5 Acknowledgments **99**

3.1 Abstract

In order to understand what expected normative behavior should be in our task we modeled reinforcement learning agents based on a recently published theoretical framework [Drugowitsch et al., 2012]. The authors of [Drugowitsch et al., 2012] used the partially observable Markov decision process framework to study optimal decisions for the standard 2AFC task. We generalized the framework to the embodied task by subtracting from the expected rewards for correct completion of the task at different times, a cost for moving towards the targets in a discrete grid world. Our RL agent approaches the target she believes to be correct, and can make correcting movements depending on the width of the arena and the urgency and movement costs, capturing the phenomenology of longer choice trajectories for more difficult trials. For a fixed movement cost, it is possible to see that time costs and arena widths have differential effects on the behavior of the agent. The time cost determines the magnitude of the beliefs on the stimulus required for the agent to move rather than wait by setting the height of her soft or reversible decision bounds, whereas the width of the arena can commit the agent to earlier choices by shifting the agent's hard irreversible decision bounds to earlier times.

3.2 Introduction

Reinforcement Learning (RL) can provide optimal solutions to hard problems, so long as the goal of the problem and all the rules of the game can be defined [Sutton and Barto, 2018]. In the past decade there has been an increased interest in the field as deep RL agents managed to beat expert humans in strategy games with millions of possible sequences of moves like

chess [Silver et al., 2018] and Go [Google, 2015, Silver et al., 2017], and even more complex real-time multi-player strategy video games such as StarCraft II [DeepMind, 2019]. Until recently in the field of decision making the most popular framework to model choices was still the classic sequential probability ratio test (SPRT) [Wald and Wolfowitz, 1948], typically in the form of the mathematically equivalent drift diffusion framework [Stone, 1960, Ratcliff et al., 2016]. Recent publications have managed to bridge insights from the SPRT and RL to find normative solutions to the binary discrimination accumulation task with time pressure [Rao, 2010, Moreno-Bote, 2010, Drugowitsch et al., 2011, Drugowitsch et al., 2012, Huang and Rao, 2013] . Drugowitsch and his colleagues extended the classic SPRT result to cases where the reliability of the evidence is unknown to the agent. The agents simulated in that study can find the optimal trade-off between certainty against the cost of time through Bayesian inference and dynamic programming [Drugowitsch et al., 2012]. We generalized this framework to consider a spatial cost for movements in the arena. In the first iterations of our model, we simplified movement costs by penalizing the velocity of changes in position. Excitingly, our model displayed both fully committed as well as hesitant trajectories, for a particular range of time and movement costs. However, this model's trajectories were series of jumps rather than smooth curves. A more accurate representation of the arm trajectories would require much more computational power, since all possible accelerations and jerks to change positions would have to be simulated in addition to all possible positions and velocities to get to them. We were not interested in understanding the physical cost of moving an arm, we were rather interested in understanding the (more cognitive) geometric cost related to the time that it takes to traverse a distance. In order to highlight this geometric aspect, we further simplified movements in the agent at each time step to a unit step to the left, right, or waiting. We modeled this in the agent as a constant cost per unit of movement. We do not explore the parameter for the cost of movement, but rather keep this number at some small fixed value and explore instead the number of movements that are required to reach the sides of the arena to report a choice. We think this is the simplest model that complicates the SPRT framework to more naturalistic choices, such as one requiring movement. We modified the task in order to reflect this new situation. We replace the joystick with button presses, but do not limit the number of presses per time step in the task. Our goal is to constraint move-

ments such that they correspond closer to the situation we are modeling, but in subjects more so than in the model the specific parameters of the arena size and time pressure have to be optimized in order to observe the hesitant state. Anecdotally, pilot experiments with very small arenas (2 or 3 required presses) did not result in hesitant trajectories whereas the model with these arena sizes continues to produce hesitant behavior. As we observed in experiment 1, in the absence of time pressure, the default behavior of subjects is not to move until they are committed to their final choice. By studying the model, we convinced ourselves that the same task with discrete rather than continuous movements (i.e. button presses) would allow us to observe the hesitant behavior, and designed that task for human subjects. The modeling results in (Figure 3.17) were useful to understand the arena sizes that maximize the occurrence of hesitant behavior. Chapter 3 describes the normative model of this task with discrete movements, and Chapter 4 describes the corresponding implementation in a human psychophysics task.

3.3 Methods

3.3.1 Task

Unlike the first two experiments, in task 3 and the following model, the subject/agent is not rewarded according to the experienced stimulus, but according to the theoretical bias of the stimulus. This is done in order to close the door to strategies that minimize integration (e.g. respond as soon as possible after counting a few frames from the stimulus). Subjects and the agent must now reach one of the two distally separated targets before the time limit, in view of temporally discounted rewards.

3.3.2 Partially Observable Markov Decision Processes and Bayesian Belief

We base our work on the Partially Observable Markov Decision Processes (POMDPs) framework described in previous studies [Rao, 2010, Moreno-Bote, 2010, Drugowitsch et al., 2012, Huang and Rao, 2013]. The POMDP framework makes the assumption that 'at any particular time step, the environment is in a particular hidden state, that is not directly accessible to the animal. This

hidden state however can be inferred by making a sequence of sensory measurements' [Huang and Rao, 2013]. The POMDP agent computes a belief, which is a posterior probability distribution over hidden states, of the world given observations. The agent uses Bayesian statistics to update her belief about the hidden state after each new observation' [Huang and Rao, 2013]. At a current belief state, an action is selected according to the learned policy, which maps belief states to actions. In the random motion task, an agent must decide the motion direction of a noisy stimulus by computing her belief over the unknown direction and motion strength from sequences of noisy observations, and selecting actions based on these beliefs to maximize the expected sum of future rewards [Huang and Rao, 2013]. At each time step, δt, the agent can choose either Left(-) or Right(+) to instantaneously report what she believes the average sign of the stimulus to be, or delay her response in order to sample more sensory evidence. The agent associates a value to each possible future state (Left, Right, Wait). This value function is computed by subtracting from all expected future rewards at each state the expected cost of reaching that state (the cost of acting now vs delaying the choice). An optimal agent with belief, g, calculates future rewards by taking the product of her Bayesian posterior beliefs, g' = g(t+1), and the reward size. This calculation involves an expectation with respect to future belief, and it is therefore necessary to compute the transition probabilities over belief states,p(g(t+1)|g(t), a), for each action. When the animal chooses to sample, at = t, the animal's belief distribution at the next time step, g', is computed by marginalizing over all possible observations [Rao, 2010]. According to the principle of optimality (see the Bellman equation in the next section), by choosing locally at the next time step the option that has the highest estimated value for all future steps, the agent maximizes her notion of long-term value. The value function is learned recursively from the known future rewards all the way to the initial state by back-propagation. Once the agent knows the action that maximizes the value function at each time step, she no longer needs the value function to act, and can enact her policy directly from her memorized map of belief states to actions. In the typical 2 choice task, which we commonly refer to as the disembodied situation, this policy is the decision bound. We adopt and generalize this model to include the simplest additional naturalistic constraint we could think of, a constant cost per unit of displacement. Our 'embodied' POMDP agent must take into consideration this cost, and critically, the time that it takes to traverse

an arena. By varying the arena size and the cost of time (aching to the arena length), we can study the behavior of an optimal agent in response to changes in the shape of the environment. We think that this consideration of geometric costs constitutes a contribution to the study of decision making by showing how these models are capable of solving more naturalistic choices. The advantage of these models are that in addition to giving a prediction that could be used to validate against data, they also provide an explanation for their behavior.

3.3.3 Bellman equation for embodied 2AFC task

The principle of optimality breaks down complex optimization problems into sequences of simpler problems. This principle is abstracted by the Bellman equation into a recursive relationship that writes the value of a state at a certain point in time in terms of the rewards making from some initial choices and the value of the future states that result from those initial choices. The following is a description and implementation of how to write the Bellman equation for a reinforcement learning (RL) agent, which discriminates a stream of stochastic evidence, and chooses one of two spatially distinct targets depending on the average sign of the stimulus. The agent moves in a manner that maximizes rewards for correct answers in light of costs for time passing and moving in the arena. Although the decision is binary and categorical, we consider space as an additional constraint, explicitly modeling graded degrees of freedom as an attempt to represent an agent physically moving in an arena. In the first iterations of our model, we simplified movement costs by penalizing the velocity of changes in position. Although the model produced hesitant decision trajectories, the model's trajectories were rather blocky or squared looking rather than the smooth curved trajectories observed in the subjects. Rather than complicating the model to produce smooth trajectories, we realized that further simplifying it to consider the same 3 choices at each time step, left, right, and waiting, would highlight the geometric aspect of having to traverse some distance, rather than the physical cost of doing it. We generalize the POMDP framework from [Drugowitsch et al., 2012] to include a unit cost per movement, and limit actions to one movement per time step. Choosing correctly results in a reward, R_{ll}, for leftward stimuli and, R_{rr}, for rightward stimuli. Incorrect responses, R_{lr} and R_{rl}, are typically assigned negative values (punishment)

or zero value (no reward). Drugowitsch and colleagues describe the value function of an agent that trades off certainty about the stimulus for earlier (potential) reward given that there is a cost for the passage of time in the task. He solves the Bellman equation with this value function to estimate the long-term estimates of reward minus cost for each possible next action. The output of this algorithm is a look-up table of next best action (the one with highest value) for each time and belief in the standard 2AFC task. The outcome is an agent that trades off speed for accuracy, similar to animals and humans performing this task. The bellman equation for the standard 2AFC task is below: With t = time, and g = belief,

$$Value(t,g) = max_{Actions}\{Expected\,Value(t+\delta t, g, action)\} =$$

$$\begin{array}{lll} & Action & Expected\,Value \\ \max\{ & \text{Right}, & gR_{rr} + (1-g)R_{lr}, \\ & \text{Left}, & (1-g)R_{ll} + gR_{lr}, \\ & \text{Wait}, & \langle V(t+\delta t, g(t+\delta t), |g,t)\rangle_{\,g(t+\delta t)} - ct\delta t \quad \} \end{array} \quad (3.1)$$

where $g(t+\delta t) \in (0,1)$, is the agent's belief or confidence that the stimulus favors the right side ($g \to 1$) or the left side ($g \to 0$), and ct is the cost per unit time δt. The agent computes the expected value of each state by summing the products of the value of rewards reachable from her current state, and the probability of collecting those rewards from that state.

Considering only the case when wrong trials are unrewarded ($R_{lr} = R_{rl} = 0$), the value function becomes

$$Value(t,g) = max_{Actions}\{Expected\,Value(t+\delta t, g, action)\} =$$

$$\begin{array}{lll} & Action & Expected\,Value \\ \max\{ & \text{Right}, & gR_{rr}, \\ & \text{Left}, & (1-g)R_{ll}\ , \\ & \text{Wait}, & \langle V(t+\delta t, g(t+\delta t))|g,t)\rangle_{g(t+\delta t)} - ct\delta t \quad \} \end{array} \quad (3.2)$$

We generalize this equation to an embodied task where the agent must navigate towards one of two targets on a 2D grid (XY) in order to collect

rewards. At each time step, the agent moves to position (x',y')$\in XY$ as to maximize her value function, which now includes an additional cost for moving. We originally incorporated a cost proportional to the squared magnitude of the agent's instantaneous velocity to reach position (x', y') in order to model experiment 2. However, in our final task we automatically displace the agent (or subject cursor) in the y direction at constant velocity towards a ceiling that corresponds to the time-limit in the task, and allow the agent to answer at any time by reaching the lateral walls of the arena. The y position when the agent reaches either side has a 1-to-1 correspondence to the reward minus time cost for that trial. This modification was made in order to make the cost in time more explicit, and simplify the control of the cursor, which now consists of discrete steps to the left or to the right. Explicitly, the goal of the task is to maximize the number of points, obtainable by reaching the right (left) boundary for positive (negative) values of stimulus bias μ as early as possible.

At each time-step δt, the agent chooses how many steps to move, paying a cost cx per unit x displaced. The agent can move to the Left or Right terminal states (the arena boundaries), remain at the same x position (wait), or move to a different x position. Each move $x \to x'$ occurs over time step δt. Waiting and moving to non-terminal x positions results in the agent experiencing one more stimulus sample, being automatically displaced by 1 unit in y, and paying the corresponding time cost $ct\delta t$. The rewritten Bellman equation is as follows:

With t =time , g =belief , δx = x displacement, cx = movement cost, $C_{\delta t}$ = time cost TR = right terminal state, and TL = left terminal state (T = terminal),

$$Value(t, g, x_t, y_t) = max_{Actions}\{ExpectedValue(t+\delta t, g, x_t, y_t, x_{t+\delta t}, y_{t+\delta t})\} =$$

$$\begin{array}{lll} & \textit{Action} & \textit{Expected Value} \\ \max\{ & \mathrm{x}_t \to x_{TR}, & \mathrm{gR}_{rr} - cx\delta x, \\ & \mathrm{x}_t \to x_{TL}, & \text{(1-g)}\mathrm{R}_{ll} - cx\delta x, \\ & \mathrm{x}_t \to x_{t+\delta t} \,\forall\, x_{t+\delta t} \notin x_T, & \langle V(t+\delta t, g(t+\delta t), x_{t+\delta t}, y_{t+\delta t}|g,t)\rangle_{g(t+\delta t)} \\ & & \text{- ct}\delta t - cx\delta x \quad \} \end{array} \tag{3.3}$$

With

$$\langle V(t+\delta t, g(t+\delta t), x_{t+\delta t}, y_{t+\delta t}|g,t)\rangle_{g(t+\delta t)} =$$

$$\int_{g(t+\delta t)} p(g(t+\delta t|g,t))V(t+\delta t, g(t+\delta t), x_{t+\delta t}, y_{t+\delta t})$$

obtained by integrating the product of the analytical solution to the belief transition probability, $p(g(t+\delta t|g,t))$, described below and $V(t+\delta t, g(t+\delta t), x_{t+\delta t}, y_{t+\delta t})$ obtained by iterative back-propagation of the value function, assumed to equal gR at the terminal state $V(T,g)$ for a T long enough that the agent must by then be sure of her belief on the stimulus.

3.3.4 Belief Transition Probabilities

In the Drugowitsch and colleagues framework, the stimulus is brownian noise, meaning "it is momentarily gaussian distributed $\delta x(t) \sim N(\mu\delta t, \delta t\sigma^2)$ for a small period of time δt , with mean $\mu\delta t$ and variance $\delta t\sigma^2$, where $|\mu|$ is the stimulus strength, and $\mu \geq 0, \mu \leq 0$ correspond to rightward and leftward favoring stimuli respectively. This evidence corresponds to a drift diffusion model $dx/dt = \mu + \eta(t)$ with drift rate μ and noise $\eta(t) \sim N(0,\sigma^2)$ and $x(t)$ describes the trajectory of a drifting/diffusing particle," [Drugowitsch et al., 2012]. In order to make sense of sensory evidence, the disembodied agent generates a posterior belief (g) on μ given available evidence $\delta s_{1:t}$, up until time t, which is the product of her prior about the stimulus $s = N(\mu,\sigma^2)$ with the likelihood of experiencing a stream of evidence $\prod_n N(\delta s_n|\mu\delta t, \delta t\sigma^2)$. This results in the following belief transition probability function:

$$p(g(t+\delta t)|g(t),t) = 1/\sqrt{\delta t_{eff}} exp($$
$$[\Phi^{-1}(g(t+\delta t))]^2/2 - (\Phi^{-1}(g(t+\delta t)) - \sqrt{(1+\delta t_{eff})}\Phi^{-1}(g(t)))^2/2\delta t_{eff}$$
) [Drugowitsch et al., 2012]

We decided to simplify the stimulus in order to reduce subjective ambiguity regarding the importance of each stimulus sample. A stimulus that can only be clearly to the left or clearly to the right for any given sample puts the focus in estimating the long-term proportion of right to left samples, and should result in easier to interpret behavior. The basic behavior of the[Drugowitsch et al., 2012] model for the standard 2AFC task is also true when we use our new stimulus, namely that psychometric functions become sharper (improved performance) and reaction times increase as the cost per time passing is lowered. We detail the new stimulus and how to turn it into an optimal belief using Bayesian inference before presenting our modeling results.

3.3.4.1 New Stimulus

We created a simpler stimulus by sampling each momentary evidence from a binomial distribution (rather than Gaussian distribution for the classical random dot visual stimulus assumed in [Drugowitsch et al., 2012], or the uniform distribution for experiment 2), such that for any one δt, the stimulus is completely lateralized to one side. The task remains to estimate the mean bias.

The new stimulus consist on a sequence of sound presentations or "bursts" of 100 milliseconds duration, each burst fully to the right or to left ($|\text{ILD}| =$ 6dB). The time between the onset of one sound presentation and the next is $\tau = 250$ ms. The full sequence is 7 seconds long, which means there are 28 total bursts. The sequence is chosen so it contains more bursts to one side than the other, in one of three different fixed proportions. Leaving costs aside, the task that the subjects need to solve consists in determining which is the side that is represented more. But because the subjects are penalized in time, they can only experience a random sample of the full sequence before committing to a decision. Therefore, the optimal solution involves inferring the probability π of the right being the more represented side. The inference problem can be idealized as the subjects experiencing Bernoulli samples of the probability π equal to the fixed proportion of right versus left in the sequence. In other words, it is equivalent to the problem

of inferring the probability of a Binomial distribution with a limited number of samples from it.

At any given time t, let us define $\bar{n}(t)$ as the total number of sound presentations up until that time. Because the interval between bursts is regular, $\bar{n}(t)$ and t play a completely equivalent role. We define $\bar{r}(t)$ and $\bar{l}(t)$ as the number of sound presentations to the right and to the left up until time t respectively. They satisfy $\bar{n} = \bar{r} + \bar{l}$ (in general we will omit the time dependence).

Let us define the number of *perceived* observations to the right as $r(t)$ and to the left as $l(t)$, both real numbers in general. Now let us assume two situations.

3.3.4.2 Noiseless counting

In this scenario, the subject is able to count perfectly $\bar{r}$ and $\bar{l}$, and therefore $r(t) = \bar{r}(t)$ and $l(t) = \bar{l}(t)$. They are integers that are incremented by 1 or 0 after each observation.

Because we are trying to infer the probability of Binomial distribution, the posterior probability is going to be a Beta distribution. Therefore is convenient to define the prior as a Beta distribution too:

$$p(\pi) = \frac{1}{B(\alpha, \beta)} \pi^{\alpha-1}(1-\pi)^{\beta-1} \tag{3.4}$$

Where the beta function $B(\alpha, \beta)$ is defined as

$$B(\alpha, \beta) = \int_0^1 \pi^{\alpha-1}(1-\pi)^{\beta-1} d\pi = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{3.5}$$

The parameters $\alpha, \beta \in R$ control the shape of the distribution. From now on we will choose $\alpha = \beta = 1$, which gives rise to a uniform prior distribution. Under these conditions, the posterior distribution of π given the observations becomes:

$$p(\pi|r, l) = \frac{1}{B(r+1, l+1)} \pi^r(1-\pi)^l \tag{3.6}$$

The belief, i.e. the probability of π being greater or equal than 0.5 is

$$g(r, l) = \frac{1}{B(r+1, l+1)} \int_{0.5}^1 \pi^r(1-\pi)^l d\pi = 1 - I_{0.5}(r+1, l+1) \tag{3.7}$$

Where $I_x(a, b)$ is the regularized incomplete Beta function.

The probability of $r'(t+\tau)$, i.e. the number of right observations after the next sound presentation is

$$\begin{aligned} p(r'(t+\tau)|r(t), l(t)) &= \int_0^1 p(\pi|r, l) \left[\pi\delta(r' - r - 1) + (1-\pi)\delta(r' - r)\right] d\pi \\ &= \frac{r+1}{n+2}\delta(r' - r - 1) + \frac{l+1}{n+2}\delta(r' - r - 1) \end{aligned} \tag{3.8}$$

Therefore the belief transition probability $g'(r', l')$ is found applying the mapping in (3.7)

$$p\left(g'(r', l')|g(r, l)\right) = \frac{r+1}{n+2}\delta\left(g' - g(r+1, l)\right) + \frac{l+1}{n+2}\delta\left(g' - g(r, l+1)\right) \tag{3.9}$$

3.3.4.3 Noisy Observations

The inference model above depends only on the probability of a right stimulus, π, on the counts experienced to the right so far, and the total counts. As is typical in decision making research, we are interested in modeling uncertainty in the inference of the bias of the stimulus. To this end, we introduce "noisy" observations, which can be flipped relatively to the stimulus with a certain probability α. This noise represents a type of sensory noise, since each stimulus sample can be occasionally counted to the wrong side. Although we take great care that each stimulus sample is clearly lateralized to one side, subjects failed to discriminate difficult stimuli perfectly, even when listening duration was extended to 12 seconds in a binary choice version of the task. In fact, performance saturated at around 80 percent for 10 second stimuli. This suggest that such sensory noise, or alternatively noise in the integration of stimuli is present in the subjects performing this task **??**. We only simulate the sensory noise since this was more straight forward to derive Bayesian equations for. The noise could be inferred for the super subject shown by the red trace in **??**, by fitting the model such that it matches the performance of the average subject for varying listening duration in the binary choice configuration. With enough sessions per subject, this parameter could be fitted for each subject. Other ways to introduce behavioral variability that were not explored include a soft-max transformation of the expected reward into action. Our choice of sensory noise results in the following discrete posterior distribution:

$$p(\pi|n,r) = \frac{1}{B(1-\alpha; r+1, n-r+1) - B(\alpha; r+1, n-r+1)} q(\pi)^r (1-q(\pi))^{n-r} \tag{3.10}$$

where n is the total number of observations and r is the number of observations to the right, $q = \alpha + \pi(1-2\alpha)$ (the "effective" probability of a right observation) and $B(z;a,b)$ is the incomplete beta function.

This leads to a belief and belief transition probability:

$$g(n,r) = \frac{I_{1-\alpha}(r+1, n-r+1) - I_{0.5}(r+1, n-r+1)}{I_{1-\alpha}(r+1, n-r+1) - I_{\alpha}(r+1, n-r+1)} \tag{3.11}$$

$$p(r'(t+\tau)|n(t),r(t)) = \left(\frac{B(r'+1,n+1-r'+1)}{B(r+1,n-r+1)}\right) \frac{I_{1-\alpha}(r'+1,n+1-r'+1) - I_{\alpha}(r'+1,n+1-r'+1)}{I_{1-\alpha}(r+1,n-r+1) - I_{\alpha}(r+1,n-r+1)} \tag{3.12}$$

where $I_x(z,w)$ is the regularised incomplete beta function and $B(a,b)$ is the beta function. The full derivation of this formula can be seen below:

The probability of an observation o_1 ($o_1 = 0$ left observation, $o_1 = 0$ right observation) given a certain value of π is:

$$\begin{aligned} p(o_1|\pi) &= \begin{cases} \pi(1-\alpha) + (1-\pi)\alpha & \text{if } o_1 = 1 \\ (1-\pi)(1-\alpha) + \pi\alpha & \text{if } o_1 = 0 \end{cases} \\ &= \begin{cases} \alpha + \pi(1-2\alpha) & \text{if } o_1 = 1 \\ 1 - (\alpha + \pi(1-2\alpha)) & \text{if } o_1 = 0 \end{cases} \\ &= q^{o_1}(1-q)^{1-o_1} \end{aligned} \tag{3.13}$$

where:

$$q := \alpha + \pi(1-2\alpha)$$

For two observations o_1 and o_2, we obtain:

$$p(o_1, o_2|q(\pi)) = [q^{o_1}(1-q)^{1-o_1}][q^{o_2}(1-q)^{1-o_2}] = q^{\sum_i o_i}(1-q)^{2-\sum_i o_i} \tag{3.14}$$

Generalizing for n observations:

$$p(o_1, o_2, ..., o_n|q(\pi)) = q^r(1-q)^{n-r} \tag{3.15}$$

$$r := \sum_i o_i$$

$$p(o_1, o_2, ..., o_n|\pi) = (\alpha + \pi(1-2\alpha))^r (1 - \alpha - \pi(1-2\alpha))^{n-r} \tag{3.16}$$

In order to calculate the posterior distribution we need to compute the integral:

$$\int_0^1 p(o_1, o_2, ..., o_n|\pi)p(\pi)d\pi = \int_0^1 (\alpha + \pi(1-2\alpha))^r (1-\alpha-\pi(1-2\alpha))^{n-r} p(\pi)d\pi \tag{3.17}$$

Using uniform distribution for $p(\pi)$ and substitution $q = \alpha + \pi(1-2\alpha)$:

$$\int_0^1 p(o_1, o_2, ..., o_n|\pi)p(\pi)d\pi = \int_\alpha^{1-\alpha} q^r(1-q)^{n-r}dq = B(1-\alpha; r+1, n-r+1) - B(\alpha; r+1, n-r+1) \tag{3.18}$$

where $B(z; a, b)$ is the incomplete beta function.

$$p(\pi|o_1, o_2, ..., o_n) = \frac{1}{B(1-\alpha; r+1, n-r+1) - B(\alpha; r+1, n-r+1)} q(\pi)^r (1-q(\pi))^{n-r} \tag{3.19}$$

The belief of the bias being to the right results from integrating the posterior for all the probabilities π that favor more right observations than left:

$$\begin{aligned} g(n,r) &= \int_{0.5}^1 \frac{1}{B(1-\alpha; r+1, n-r+1) - B(\alpha; r+1, n-r+1)} (\alpha + \pi(1-2\alpha))^r (1-\alpha-\pi(1-2\alpha))^{n-r} d\pi = \\ &= \frac{1}{B(1-\alpha; r+1, n-r+1) - B(\alpha; r+1, n-r+1)} \int_{0.5}^{1-\alpha} q^r (1-q)^{n-r} dq \\ &= \frac{B(1-\alpha; r+1, n-r+1) - B(0.5; r+1, n-r+1)}{B(1-\alpha; r+1, n-r+1) - B(\alpha; r+1, n-r+1)} \\ &= \frac{I_{1-\alpha}(r+1, n-r+1) - I_{0.5}(r+1, n-r+1)}{I_{1-\alpha}(r+1, n-r+1) - I_\alpha(r+1, n-r+1)} \end{aligned} \tag{3.20}$$

where $I_x(z, w)$ is the regularised incomplete beta function.

Finally, the belief transition probability is therefore:

$$\begin{aligned} p(r'(t+\tau)|n(t), r(t)) &= \begin{cases} \int_0^1 p(q(\pi)|n, r) \cdot q(\pi)d\pi & \text{if } r' = r+1 \\ \int_0^1 p(q(\pi)|n, r) \cdot (1-q(\pi))d\pi & \text{if } r' = r \end{cases} \\ &= \int_\alpha^{1-\alpha} p(q|n, r) \cdot q^{r'-r}(1-q)^{1-r'+r} dq \\ &= \frac{1}{B(1-\alpha; r+1, n-r+1) - B(\alpha; r+1, n-r+1)} \int_\alpha^{1-\alpha} q^{r'}(1-q)^{n+1-r'} dq \\ &= \frac{B(1-\alpha; r'+1, n+1-r'+1) - B(\alpha; r'+1, n+1-r'+1)}{B(1-\alpha; r+1, n-r+1) - B(\alpha; r+1, n-r+1)} \\ &= \left(\frac{B(r'+1, n+1-r'+1)}{B(r+1, n-r+1)}\right) \frac{I_{1-\alpha}(r'+1, n+1-r'+1) - I_\alpha(r'+1, n+1-r'+1)}{I_{1-\alpha}(r+1, n-r+1) - I_\alpha(r+1, n-r+1)} \end{aligned} \tag{3.21}$$

The belief transition probabilities after sampling 5 stimuli compared to after sampling 15 stimuli are shown below:

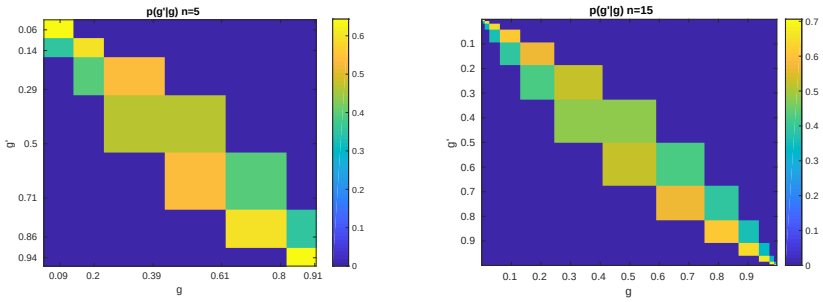


Figure 3.1: Belief transition probabilities after 5 samples of stimulus (left) and after 15 samples (right). Since the values of belief and probabilities are discrete and not equally spaced, the "squares" in the plot were scaled to approximately represent the correct distances in belief space. (the lack of the symmetry that should exist around 0.5 on both axis is a result of approximating). Colorbar shows the probability p(g'|g).

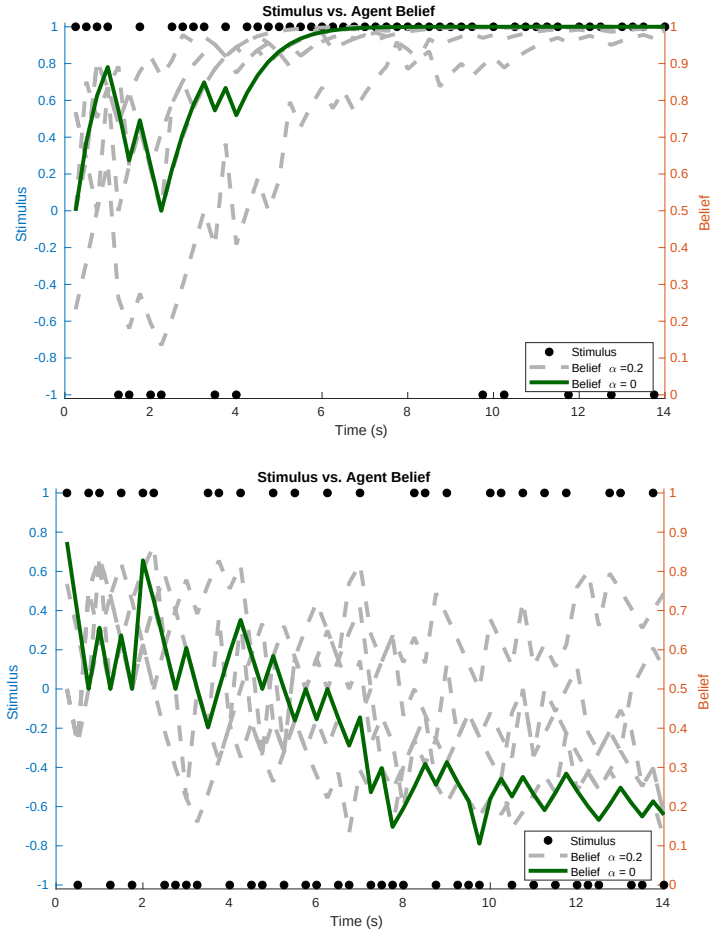


Figure 3.2: New stimulus based on a binomial distribution of ILDs. Black dots represent the stimulus. The green line is an optimal agent's belief that it is experiencing a right-sided stimulus, $p(\pi \geq 0.5)$, given the evidence until that time. Different difficulties can be created by changing the probability, π, of the stimulus being on the right, where π values less (greater) than 0.5 have a left (right) bias. The green line represents the belief resulting from perfect observations (noiseless) and the dashed grey lines correspond to the belief of 5 different trials with the same stimulus, but with noisy observations (stimulus can be "flipped" with probability $\alpha = 0.2$). After noticing that both easy (top) and a hard (bottom) stimuli can be generated with the same sampling probability of being right (0.6), we decided to sample stimuli for the model from this distribution and its left counterpart.

The new stimulus is shown above, with the resulting Bayesian belief that the stimulus favors the right side for 5 repeats of the same stimulus with noiseless inference (3.7) vs noisy inference (3.11). The stimulus is sampled with a probability of being right of 0.6. For each agent in the modeling work, 10,000 stimuli were generated, half with probability of being right of 0.6 (right favoring stimuli) and half with probability 0.4 (left favoring stimuli), fixing the stimulus noise parameter at $\alpha = 0.2$. A wide range of easy and hard stimuli result from this sampling. In order to label the side and difficulty for a particular trial, we count the actual sampled proportion of left to right samples with respect to the total number of samples,

$$(N_R - N_L)/(N_R + N_L),$$

(3.22)

and we use the absolute value of this quantity as measure of difficulty for the trial independent of side,

$$|(N_R - N_L)/(N_R + N_L)|. \tag{3.23}$$

For all the results detailed in this section, the value function is always back-propagated from T=14 seconds, which is much longer than the average reaction times (4 seconds) we are interested in studying. The urgency of the agent is set with the cost of time. We commonly refer to the concept of time limit in the analysis. Time limit corresponds to how many times the total duration of a trial (T) would mean a loss of all possible reward points. For example, a time limit of 1 means that all possible reward points are lost at the end of a trial. On the other hand, if the time limit is 4, all possible reward points would only be lost if the reaction time was four times the duration of a trial (which is not possible), which means that one fourth of the points are lost at the end of a trial. Notation: **y#** means a time limit of #.

In order to understand the behavior of our embodied agents relative to the width of the arena, the size of the arena is varied while a unit cost per action is fixed at 0.01 for all experiments.

3.4 Results

3.4.1 Normative model of the disembodied 2AFC task

We first replicate the main results of [Drugowitsch et al., 2012] for the disembodied task (arena size 3, meaning left, wait, and right states) corresponding to the classical 2AFC discrimination task, using our own stimulus. Below we show a disembodied agent that loses all the possible points linearly by 4 seconds, back-propagating the value function from T = 14 seconds. All figures are labeled according to the time limit by which any positive reward would be lost given some particular cost per time step ($\delta t = 0.25$ seconds). Reaction times decrease as a function of the stimulus strength (3.23).

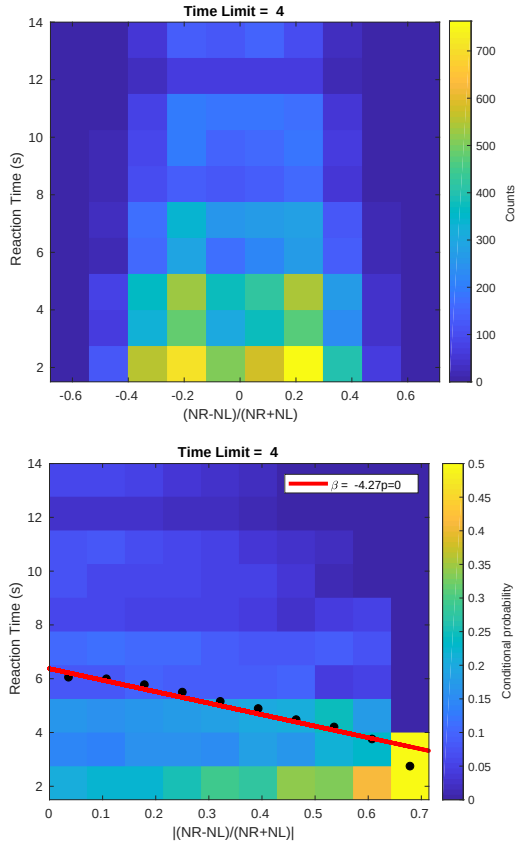


Figure 3.3: Reaction times of a normative model of the standard 2AFC task with our stimulus. Top panel: Stronger biases to the left (right) are represented by more negative (positive) values on the horizontal axis. Bottom panel: left and right trials are collapsed by taking the absolute value of the difference of the number of right (NR) and left (NL) counts relative to the total number of counts), the stimulus strength (3.23). Colorbar represents the conditional probability of the reaction times given a certain stimulus strength. Black dots represent the mean reaction time for each bin of stimulus strength. Red line is the linear regression of Reaction time onto the stimulus strength.

The increase in reaction time with increasing difficulty gets steeper when difficulty is defined as the absolute mean of only the experienced stimulus.

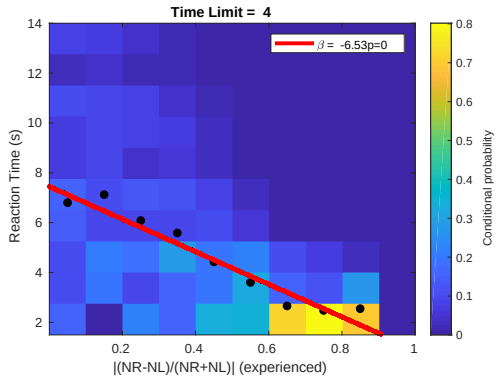


Figure 3.4: Reaction time as a function of the absolute experienced stimulus strength. Colorbar represents the conditional probability of the reaction times given a certain experienced stimulus strength. Black dots represent the mean reaction time for each bin of experienced stimulus strength. Red line is the linear regression of Reaction time onto the experienced stimulus strength.

As in the Drugowitsch and Moreno frameworks [Drugowitsch et al., 2012], when the cost in time is relaxed, reaction times increase and the agent experiences more evidence, which results in better performance.

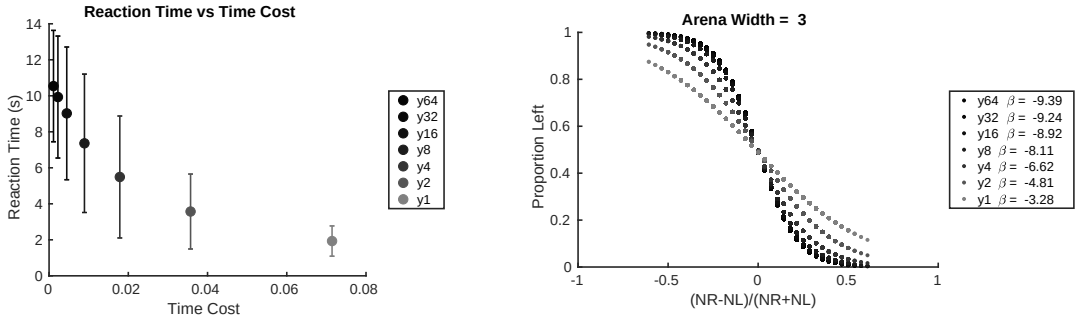


Figure 3.5: Normative behavior for a single press 2AFC task (arena size 3, meaning left, right and wait states). Left panel: Reaction times increase as the cost of time is decreased (equivalent to increasing the time limit), error bars denote standard deviations. Right panel: Psychometric curves sharpen with decreased cost of time. On both panels, the legend shows the Time Cost converted to the equivalent time limit y (which is proportional to the time at which a positive score is no longer possible). The right panel legend additionally shows the slope parameter for the psychometric curve for each Time Cost.

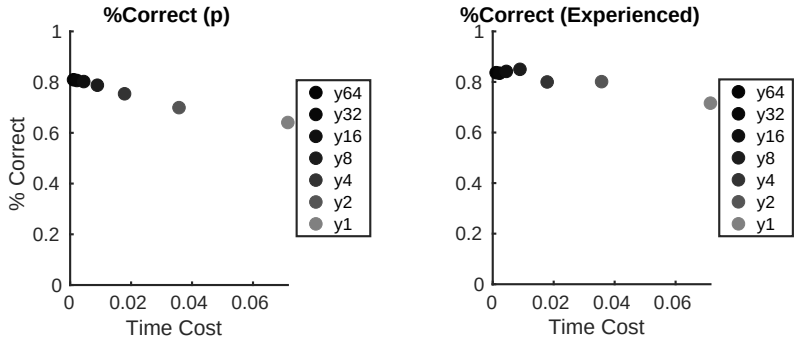


Figure 3.6: Proportion of correct responses increases with decreasing time cost (increasing time limit) whether performance is quantified as a function of the long term mean of the stimulus, p (left panel), or as a function of the experienced stimulus (right panel). On both panels, the legend shows the Time Cost converted to the equivalent time limit y (which is proportional to the time at which a positive score is no longer possible).

Since in our first two experiments we rewarded subjects as a function of the stimulus they experienced, rather than on the actual long-term mean of the stimulus, we also computed the performance of the agent according to the experienced stimulus. Two main observations can be made. First, for all time limits performance would be higher if we scored according to the experienced stimulus up to the time the agent reports her choice. This suggests that the agent always thinks she is doing better than she actually is when she answers. Beliefs higher than average might lead to decisions whereas lower beliefs would lead to waiting, leading to overconfidence. Second, the pattern of improved performance for longer time limits can also be observed when quantifying performance in this manner, although the results are noisy. For this reason we score agent behavior as a function of the long-term mean of the stimulus (condition), and in our final experiment, described in chapter 4, we reward subjects according to whether the long-term mean of the stimulus is positive or negative.

At each time point, the belief for which the expected long-term value of waiting intersects the expected value of choosing left (right) defines the left (right) decision boundary. In other words, for each time there exists a belief for which waiting longer will never result in a higher reward regardless of how much more certain the agent becomes by doing so, so the agent should choose according to her present belief at that time.

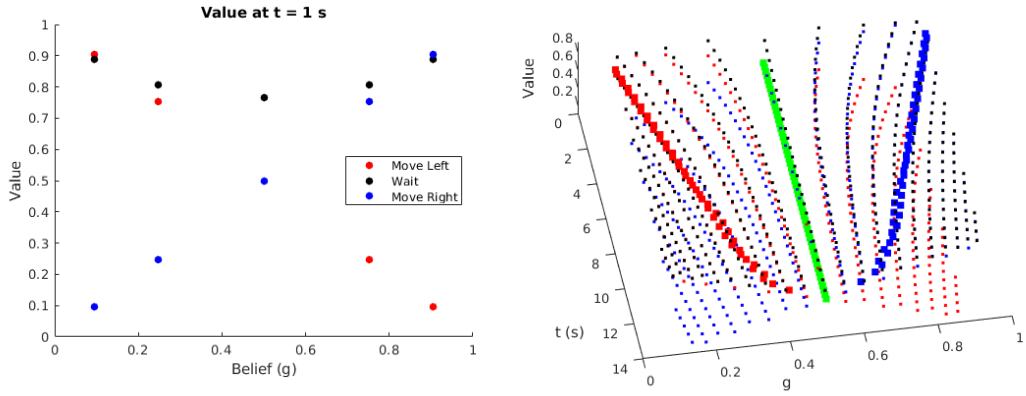


Figure 3.7: Value as a function of the agent's belief. To the left, a slice at t=1 s is shown, representing the values of waiting (black), choosing left (red) or choosing right (blue) for all possible beliefs. To the right, the value function is shown for all times, with larger dots representing the intersection at each time of the curves of waiting with the ones of moving left (large red dots) and the curves of waiting with the ones of moving right (large blue dots). These points correspond to the decision boundaries, where the agent should move left (right) if belief is smaller (bigger), than the one of the decision boundary. Large green dots correspond to the intersection of the value curves for moving left and moving right.

These bounds determine the choices for a given stimulus belief, such that an agent with zero noise in the integration of the stimulus will always have the same beliefs for repeats of the stimulus, and thus always make the same choice. However, variance in responses to a fixed stimulus is introduced by the noise parameter α.

As previously described [Drugowitsch et al., 2012], decision boundaries on the agent's belief collapse later as the cost of time decreases (time limit, y, increases), which increases reaction times and improves performance.

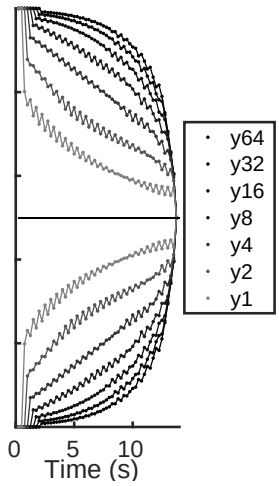


Figure 3.8: Time evolution of decision boundaries in belief space for different time limits. The decision boundaries collapse in time. The distance between the bounds widens as the cost of time is decreased (time limit, y, is increased). Each pair of mirrored lines represents the threshold or decision bound that the agent's belief must cross at each time in order to make a choice for a particular time limit imposed by a cost for time passing. The legend shows the Time Cost converted to the equivalent time limit y (which is proportional to the time at which a positive score is no longer possible).

3.4.2 Normative model of our embodied 2AFC task, fixing arena width and time limit parameters

We next characterize the behavior of a normative agent that must make several moves in order to report a choice, and then show what happens to the bounds as we move from the disembodied task to the embodied task. As before, we first show behavior for a fixed time limit of 4, and then for a range of time costs as before, except now the number of presses necessary to reach either side is fixed at 8 (arena size = 2*8 (sides) + 1 (center) = 17). Left and right stimuli have been collapsed using (3.23) so that higher values mean easier stimuli. As with the disembodied agent from Figure 4.2, reaction times decrease as the difficulty of the stimulus decreases (higher absolute mean ILD). As the difficulty of the stimulus increases, the agent also uses a significantly higher number of presses to report its choice.

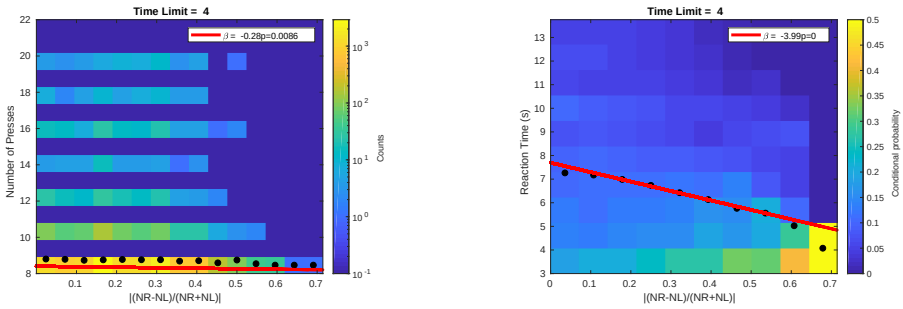


Figure 3.9: Left panel: Number of presses as a function of stimulus strength, for an arena width of 17. The colour scale represents the number of counts and is logarithmic. Black dots represent the mean number of presses for each bin of stimulus strength. Red line is the linear regression of Number of Presses onto Stimulus Strength. Right panel: Reaction times as a function of stimulus strength for an arena width of 17. The colour scale represents the conditional probability of the reaction times given a certain stimulus strength. Black dots represent the mean reaction time for each bin of stimulus strength. Red line is the linear regression of Reaction time onto Stimulus Strength.

While the decrease in mean RT appears to be more robust (larger coefficient), there is a significant dependence of the total number or presses on the difficulty of the task. It is clear from the figure that the easiest stimuli get resolved with the minimum number of presses, whereas harder stimuli can elicit more than twice that.

Remarkably for our investigation, the raw agent behavior in the arena for easy versus a harder stimulus can be qualitatively similar to the phenomenology we observed in experiments 1 and 2.

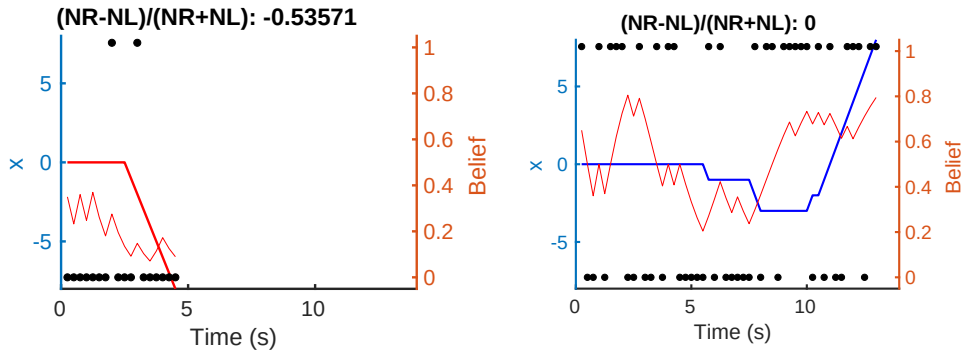


Figure 3.10: Normative agent trajectories for an arena of width 17, back-propagation from 14 seconds, unit cost per action 0.01 with stimulus noise $\alpha = 0.2$ for an easy (left) and a hard (right) trial. Black dots on the borders of each plot show the stimulus, and thin red line the corresponding Bayesian belief. A thick red (blue) line on the left (right) plot shows the decision trajectory of the agent for one presentation of that stimulus.

3.4.3 Decision bounds for the embodied normative agent reflect the agent's commitment to a choice

We begin by showing the action value intersections defining the decision bounds for the disembodied agent of (Figure 3.7), but for the embodied agent with arena width x17. The embodied action bounds to move one step to the right or left from x = 0 are more narrow relative to the disembodied decision bounds to choose left or right, however, whereas the disembodied agent reports her final choice after taking one action to either side, the embodied agent encounters another decision bound. These new decision bounds have the interesting property that the mid-point belief g = 0.5 is closer to the bound on the side of the direction that the agent moved. That is, it takes less extreme values of belief to the left (right) to make an action to the left (right) for an agent that is on the left (right) side of the arena. The bounds change with the agent's position, making the left (right) actions more valuable as the agent moves left (right), resulting in a positive feedback loop that eventually commits the agent to choose left (right) when the value of the waiting state no longer surpasses that of either bound. This occurs when the left and right bound touch or collapse. At this time, the agent is forced to commit to whatever side she is closest too. The collapse of

the bounds can be seen in the figure below, at the time-step where only a green dot but no red or blue dots (the bounds) are present anymore. The time step immediately before that is last step at which the agent still has time to make it to either side before her points run out. We call this final set of bounds denoting non-reversible actions, hard bounds, and they have the same categorical nature of the disembodied bounds that commits agents to their choices. In contrast, bounds encountered before this time step, at which the agent can still delay her final choice to either side by one more time-step, we call soft bounds. The difference between soft and hard bounds is by definition, since the bounds are derived directly from the values of each potential action (left, right, wait). Crossing a soft bound in general leads the agent to encounter a new set of bounds, however, the hard bound is the bound collapse, which is the moment for which there remains only sufficient time to reach the closest target. For an agent in the center, the hard bound collapses at 0.5, so the agent will immediately begin to move to the side that her belief favors. Although it is difficult to conceive an analysis that finds these bounds from behavior rather than action values, the consequence of the hard bounds is an agent that at some point must commit to one side or the other. This inspired the final analysis of the thesis, where we quantify probabilities of moving left, waiting, or moving right as a function of the stimulus evidence, for different times and locations in the arena 4.31.

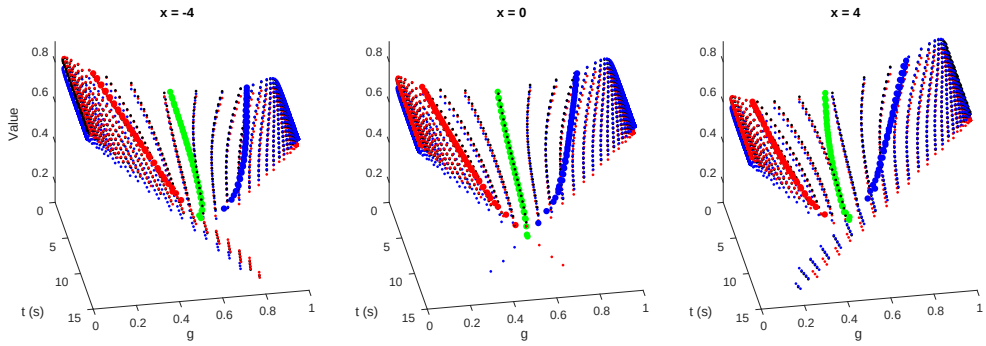


Figure 3.11: Value function for all times and for current positions (a)x=-4, (b)x=0, (c)x=4, for an arena width of 17, cost of movement 0.01, cost of time 0.018 and back-propagation from 14 seconds. Small red, blue and black dots represent the values of moving left, moving right and waiting at the current position, respectively. The larger dots represent the intersection at each time of the curves of waiting with the ones of moving left (large red dots) and the curves waiting with the ones of moving right (large blue dots). Large green dots correspond to the intersection of the curves of moving left and moving right. When the agent's belief hits the left (right) bound at the initial position, x = 0, she moves to x = -1 (x=1). Note the value of waiting at x = -4 (x = 4) marked by the green dots is closer to the red (blue) dots denoting the left (right) bound.

3.4.4 Reliability: Trajectories for repeated presentations of the same stimulus

The trajectories of the agent are fully determined by the bounds and her beliefs. This means that for noise-less inference on the integrated stimulus ($\alpha = 0$), repeating the same stimulus results in the exact same trajectory output by the model. However, noise in the memory of the integrator causes slight differences in the beliefs of the agent, causing slight differences in the output trajectories when beliefs are allowed to vary with α on a trial by trial basis. This results in qualitatively similar trajectories for repeated presentations of the exact same stimulus. Below we show the responses to repeated presentations of easy stimuli (Figure 3.12) and for hard stimuli (Figure 3.13) with and without noise.

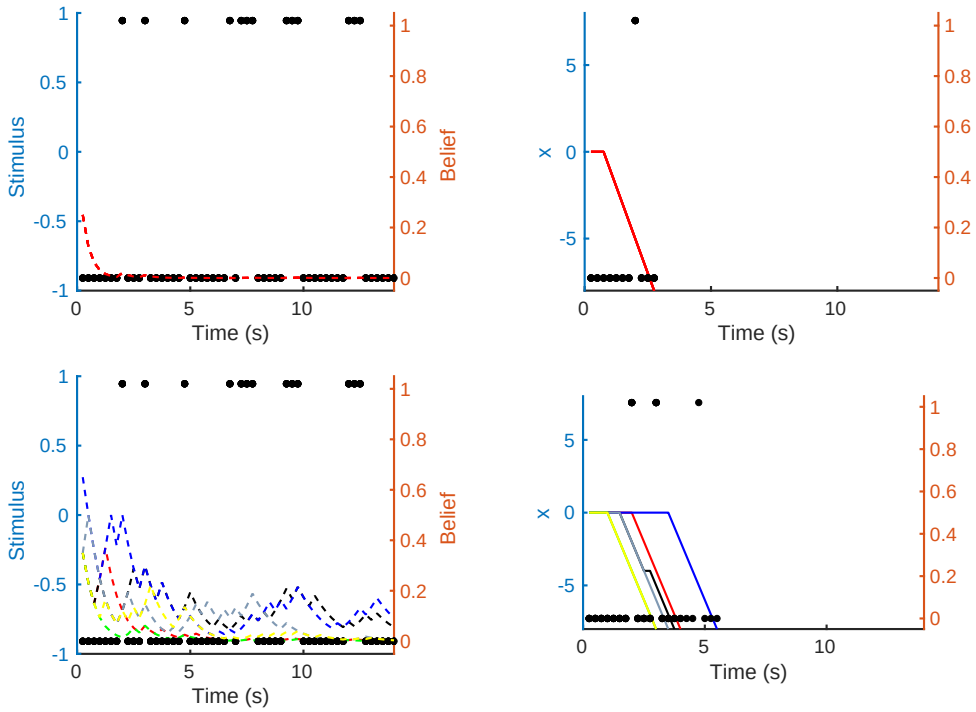


Figure 3.12: Trajectories for 5 repetitions of an easy stimulus and the resulting beliefs (left panels) and trajectories (right panels). Easy stimuli result in identical trajectories for both noiseless (top) and noisy (bottom) integration. Black dots on the borders of each plot show the stimulus. On the left plots, dashed colored lines show the Bayesian beliefs for repeated presentations of the stimulus. Thick colored lines on the right side plot shows the decision trajectories of the agent elicited by the beliefs on the left plots.

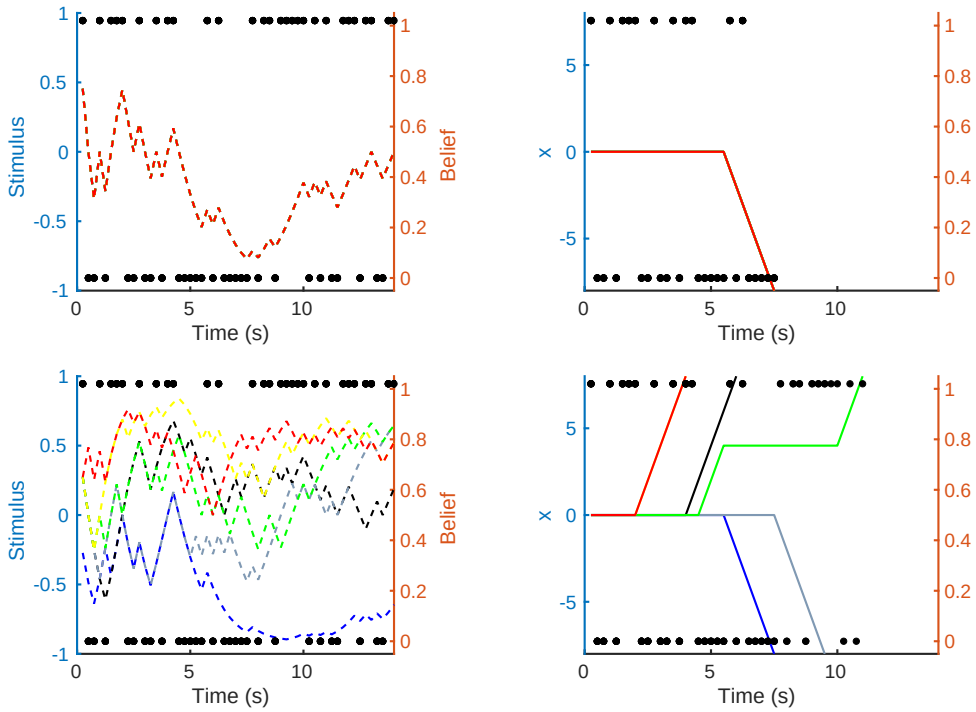


Figure 3.13: Trajectories for 5 repetitions of a hard stimulus and the resulting beliefs (left panels) and trajectories (right panels). For noiseless integration trajectories are identical (top), however with noisy integration (bottom panels) some variability in the output trajectory is introduced the more variable the early part of the stimulus. Black dots on the borders of each plot show the stimulus. On the left plots, dashed colored lines show the Bayesian beliefs for repeated presentations of the stimulus. Thick colored lines on the right side plot shows the decision trajectories of the agent elicited by the beliefs on the left plots. Trajectories remain strikingly qualitatively similar.

The beliefs and trajectories are shown together for repeats of a stimulus on a single plot below.

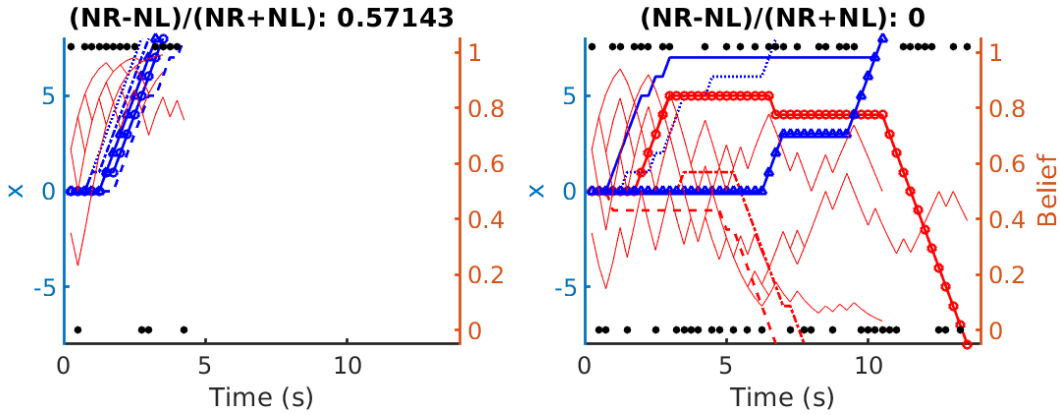


Figure 3.14: Normative agent trajectories for an arena of width 17, time limit 4, unit cost per action 0.01 with stimulus noise $\alpha = 0.2$ for 5 repeats of an easy stimulus (left) and a hard stimulus (right).The red thin lines represent the evolution of belief in the different trials, while blue lines of varying markers denote trajectories that end in the right (figure top) side. Red color likewise denotes trajectories that ended on the left.

3.4.5 Varying Time Cost with Fixed Arena Width

For a fixed arena width of 17, the influence of varying costs of time was tested. Reaction times are not as useful to describe the behavior of embodied agents, as they are not revealing of the performance of the agent or the shape of her choice trajectories. We define hesitation as the number of presses used by the agent divided by the minimum number of presses required to report a choice, and found this to be a better indicator of the complexity of the decision trajectories. We also define other observable signatures of the behavior such as start-time (the time at which the first move is made) and of Post-start waits (the number of times that the agent waits instead of moving after her first move). All of these quantities appear to change monotonically as time cost is increased, except for Post-start waiting, which is small for small time costs (large arenas), then increases as time cost is increased (arena size decreased), and then decreases for very large time costs (very small arenas) (Figure 3.17). This non-monoticity is the first indication that in the model, hesitation or related behaviors might be only observed for very particular regimes of cost. This will become clearer in the analysis by arena width (Figure 3.22).

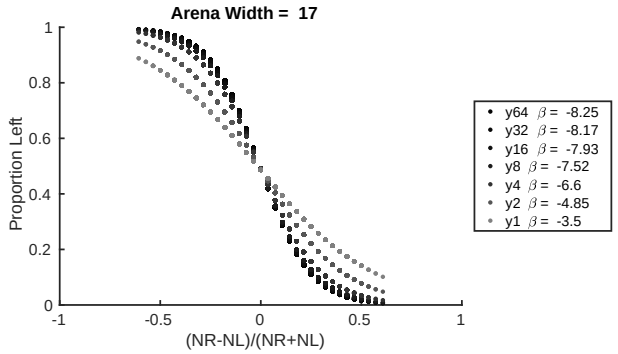


Figure 3.15: Psychometric curves for different costs of time/time limits. The legend shows the Time Cost converted to the equivalent time limit y (which is proportional to the time at which a positive score is no longer possible), and the slope parameter for the psychometric curve for each time limit.

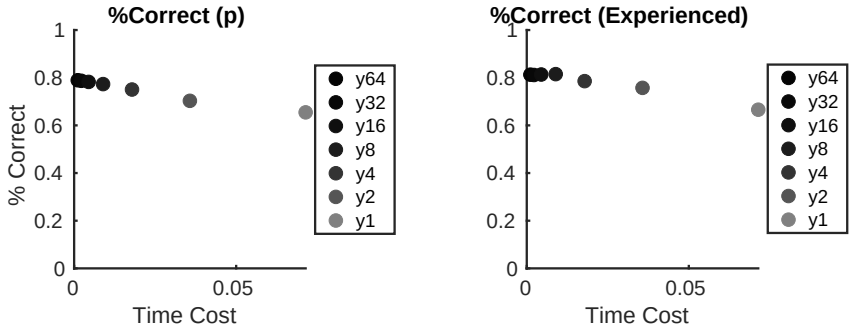


Figure 3.16: Proportion of correct responses as a function of the time cost. In the left panel, the accuracy of responses is evaluated according to the whole stimulus, while in the right panel it is evaluated according to the section of the stimulus experienced by the agent until the moment of their final choice. The legend shows the Time Cost converted to the equivalent time limit y (which is proportional to the time at which a positive score is no longer possible).

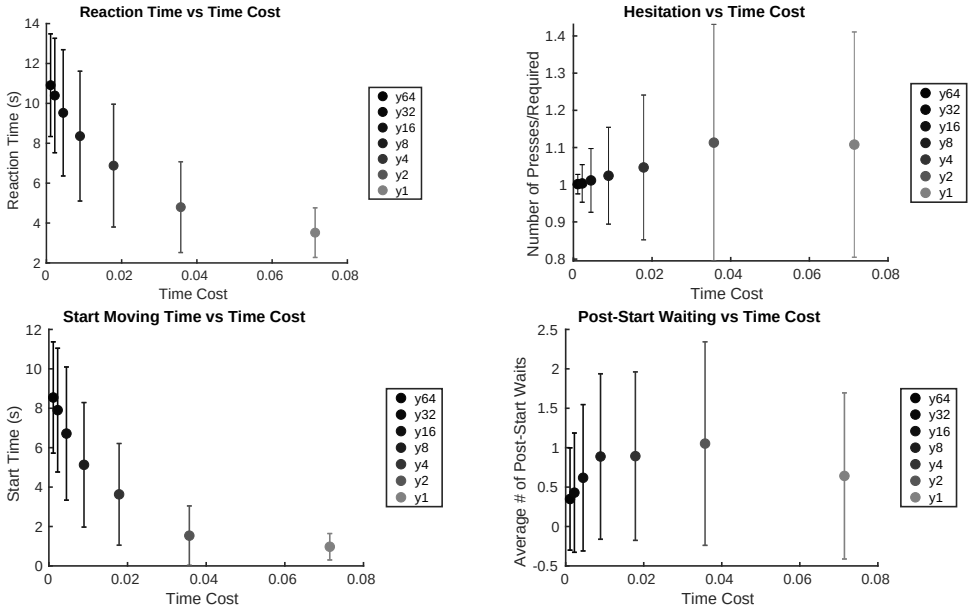


Figure 3.17: Signatures of the behaviour of the embodied agent for fixed arena width of 17 and $\alpha = 0.2$ and varying time costs/limits. Error bars show standard deviations. The legend shows the Time Cost converted to the equivalent time limit y (which is proportional to the time at which a positive score is no longer possible)

We next explore whether the bias to the bound as a function of the agent's position that we saw in (Figure 3.7) is general across different time cost regimes (time limits). We find that indeed the bounds change with the agent's position, making the left (right) actions more valuable as the agent moves left (right), resulting in a positive feedback loop that eventually commits the agent to choose left (right). Increasing time pressure (reducing the time limit) on the bounds has a shrinking effect on the bounds, making them less wide at each time-step.

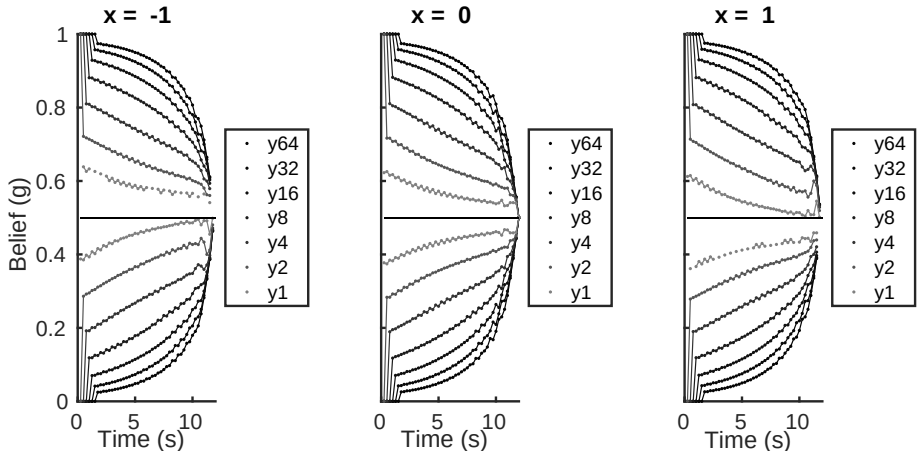


Figure 3.18: The left (right) panel shows the action bounds to move one step left or right for all times at position x = -1 (x = 1). Color code denotes the bounds for different time limits, for an arena of size x = 17 and stimulus noise parameter $\alpha = 0.2$. As the time limit increases, the bounds get wider. Note that for all time limits, the beliefs necessary to move left (right) indicated by the bottom (top) bound are closer (further) to a fully ambiguous belief of 0.5. Previous movements to the left (right) of the center of the arena makes the left (right) actions more valuable, resulting in a positive feedback loop that biases and eventually commits the agent to the left (right). The legend shows the Time Cost converted to the equivalent time limit y (which is proportional to the time at which a positive score is no longer possible)

3.4.6 Performance of the embodied vs disembodied agent for fixed arena width and varying time limit parameters

We computed psychometric curves to assess the performance of the normative agent for the same arena width, while varying the time limit (arena length y). We find that for the same time costs (time limits), the embodied agent has and the disembodied agent have overall similar psychometric curves, with perhaps sharper slopes for the embodied agent for smaller time limits, but the opposite is true for larger time limits. Performance as a function of arena width and time limit (cost) will be analyzed systematically in (Figure 3.24).

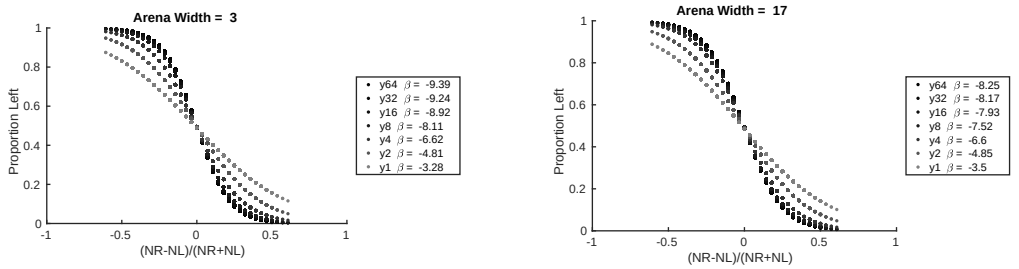


Figure 3.19: Psychometric curves for a disembodied agent (left panel) vs an embodied agent (right panel) for the same costs of time (time limits y). The legend shows the Time Cost converted to the equivalent time limit y (which is proportional to the time at which a positive score is no longer possible), and the slope parameter for the psychometric curve for each time limit.

3.4.7 Embodied normative agent behavior for varying arena widths and fixed time limit parameters

In order to quantify the effect of arena width on the behaviour of the agent, the cost of time was fixed at 0.018 (equivalent to 14 seconds time limit) and the width of the arena was varied (e.g. the minimum number of presses required to report a decision). We found that small arenas allow agents to commit later relative to the disembodied task, and make corrections, giving rise to hesitant choice trajectories. Larger arena widths force the agent to begin moving earlier and commit to wrong choices, causing a drop in performance observed in the psychometric curves.

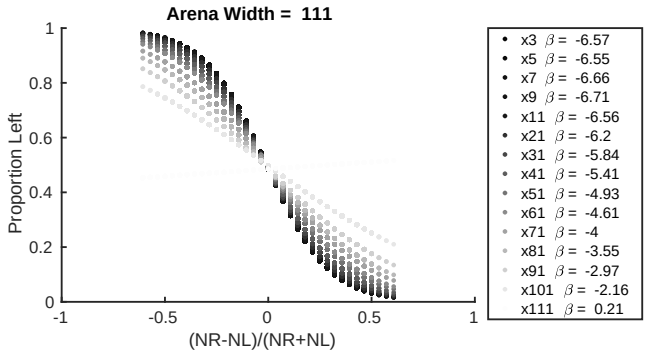


Figure 3.20: Performance for our task going from the standard single press 2AFC task (x3) to two press (x5) and larger arenas. Small embodied arenas have similar performance as the standard 2AFC task, with perhaps a small improvement for size (x7) and (x9) but for larger arenas performance drops sharply. The legend shows the arena width, and the slope parameter for the psychometric curve for each width.

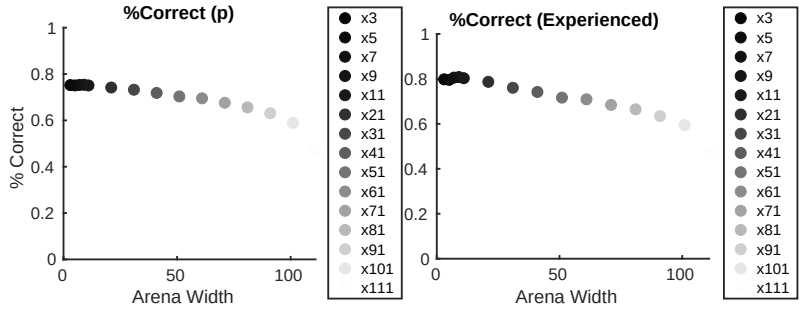


Figure 3.21: Proportion of correct responses as a function of the arena width with a fixed time limit of 4 (time cost = 0.018). In the left panel, the accuracy of responses is evaluated according to the whole stimulus, while in the right panel it is evaluated according to the section of the stimulus experienced by the agent until the moment of their final choice. For both panels, the legend shows the arena width.

We plot the signatures of behavior (Reaction time, hesitation, start moving time and post-start waiting vs the different time costs) (Figure 3.22) for the agents of (Figure 3.20), with fixed time limit of 14 with noise parameter

$\alpha = 0.2$ and varying arena widths.

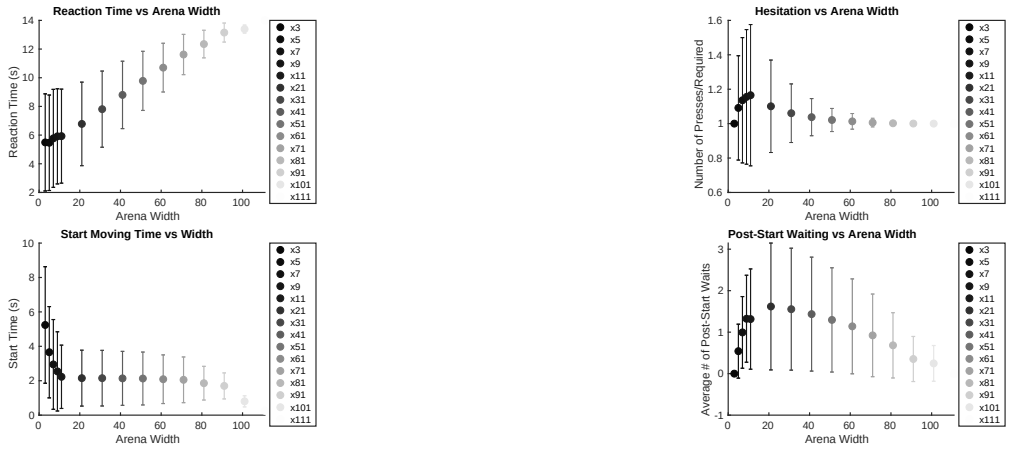


Figure 3.22: Signatures of the behavior of the embodied agents from (Figure 3.20) for fixed time limit of 14 and $\alpha = 0.2$ and varying arena widths. Error bars show standard deviations. For all panels, the legend shows the arena width.

As the size of the arena grows from the disembodied to the smallest embodied arenas (x5, x7), Reaction Times don't change much, but Start Time decreases and Hesitation (number of presses/required) increases as does Post-Start Waiting. These changes in behavior are accompanied by an increase in the percentage of correct responses relative to the experienced stimulus. Further increasing the width of the arena causes Start Times to drop, but then it remains the same for a large range of arena widths (x9-x91). Similarly, Post Start waiting increases for the next two arena sizes (x9-x11), but then plateaus for a large range of arenas (x11-x61). Percentage of correct responses and the slope of the psychometric curve also show a minor improvement in performance for (x7-x9) arena widths relative to smaller arenas, but then plateau and remain unchanged for a large range of arenas(x11-x61). Hesitation also increases at first with further increases in arena width, and reaches a peak for arena size x11, and then decays smoothly as we further increase arena width. Very large arenas (>x61) cause a drop in Post-Start Waiting and percentage of correct responses as well as a flattening of the psychometric curve. The largest arena sizes (x101-x111) push Start Times to δt, Post-Start waiting to zero, and Hesitation disappears (Number of Presses/required = 1), and performance drops to chance level. This indi-

cates that the arena is so large that given the time limit, the agent doesn't have time to integrate any evidence and acts on her first guess (Figure 3.22). This non-monotony of the hesitation behaviors, best appreciated in Number of Presses/required and in Average of Post-Start waits, is theoretical evidence that the hesitation requires very specific environmental conditions to occur.

Finally we observe what happens to the decision bounds as we increase the width of the arena. Whereas increasing arena length (increasing time limit) had a widening effect on the bounds (Figure 3.18), increasing arena width causes the bounds to collapse earlier while keeping the same width.

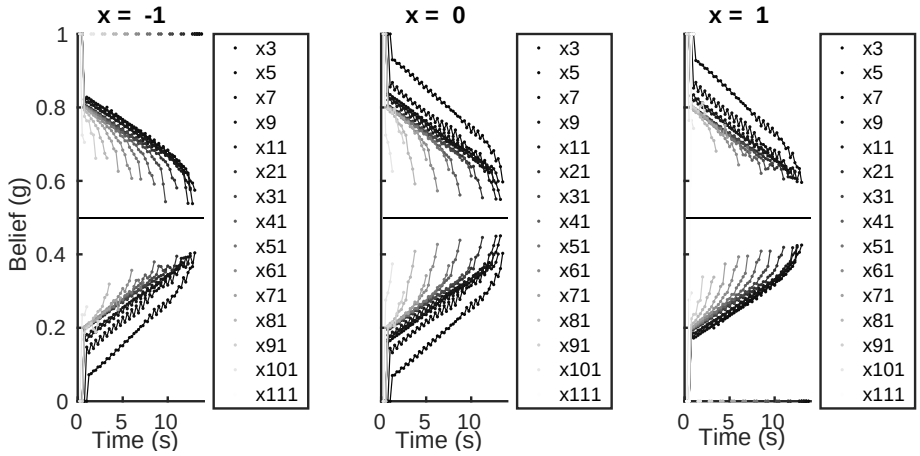


Figure 3.23: The left (right) panel shows the action bounds to move one step left or right for all times at position x = -1 (x = 1). Color code denotes the bounds for different arena widths, for fixed time limit 14 and stimulus noise parameter $\alpha = 0.2$. As the arena width increases, the bounds collapse earlier while maintaining their width. Note that as observed in (Figure 3.18), for all time limits, the beliefs necessary to move left (right) indicated by the bottom (top) bound are closer (further) to a fully ambiguous belief of 0.5. Previous movements to the left (right) of the center of the arena makes the left (right) actions more valuable, resulting in a positive feedback loop that biases and eventually commits the agent to the left (right). For all panels, the legend shows the arena width.

In (Figure 3.18) we learned that increasing the cost of time (decreasing time limit) causes the soft bounds to narrow, such that less extreme beliefs are required for the agent to move rather than wait. Meanwhile, increasing the width of the arena shifts the hard bounds or bound collapse to earlier

times (Figure 3.23). Finally, (Figure 3.12 and 3.13) showed that the agent's actions are deterministic at each position and time in the arena with respect to fixed beliefs, but noise in the beliefs introduces variability in behavior.

3.4.8 Embodied normative agent behavior for varying arena widths and varying time limit parameters

We next confirm that the results described in (Figure 3.22) generalize to other time limits. In general, as the time limit (y) increases we also observe increasing start times and the percentage of correct responses increases as well. This behavior is similar to that of the disembodied agent. Our novel finding is that small arenas allow the agent to delay her choice by moving to non-terminal states in addition to waiting. The hesitation behavior characterized by Number of presses / required presses is strongest for small arena widths and time limits. In our human experiments conducted in parallel with this work, pilot trials were used to decide on a time limit y of 7 seconds and arena width x17. This arena size is part of the maximum Hesitation region in the model. As arena sizes grow sufficiently large, hesitation behavior and Post-Start waiting disappear, Start times drop to δt and performance drops to chance level for all time limits simulated.

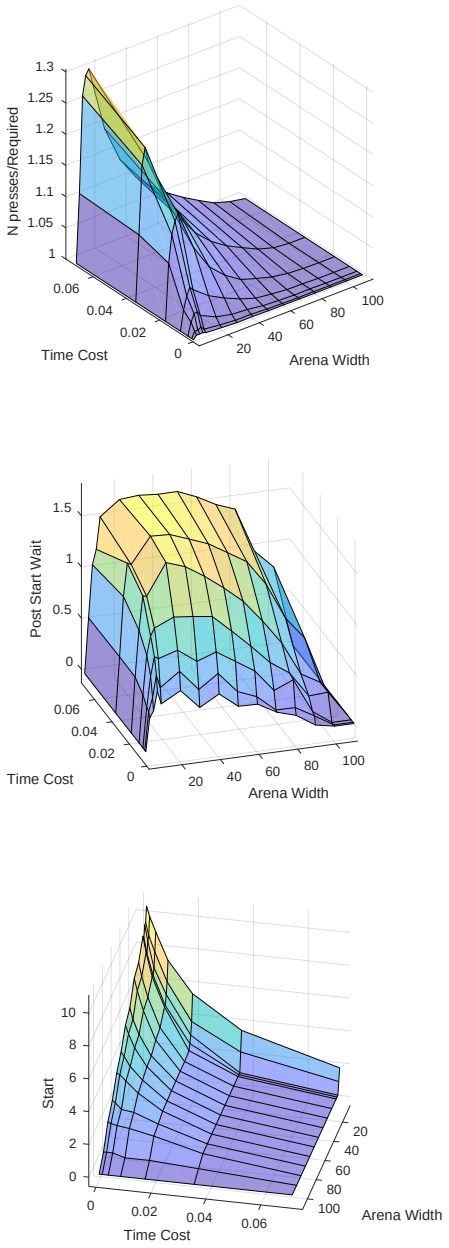


Figure 3.24: Signatures of the behavior for a range of arena widths, time limits, and noise parameter $\alpha = 0.2$. The color code is in the scale of each plot's y-limits.

While the effects of increasing arena width (X) at fixed cost of time (T) and the effects of increasing T at fixed arena width appear equivalent at the level of the psychometric (decreased performance), these two parameters have differential effects at the level of the decision bounds, which ultimately determine behavior. Although it may be possible mathematically relate one summary statistic of behavior such as percentage of correct responses to some relation between X and T, in general X and T have different effects across different signatures of behavior.

Interestingly, the signatures of behavior we think of as being most indicative of a lack of commitment to an initial choice (Hesitation = N presses/Required and Post-start waiting) are non-monotonic. The justification for the trends of both signatures of behavior should be the same, but they present different shapes of curves (post-start waiting increases faster as time pressure increases from low values than hesitation). This is due to the costs of movement, which justify that the agent, when time pressured not to do only direct trajectories, should prioritize waiting rather than moving hesitantly to avoid the costs of movement. The low values of hesitation and post-start waiting for both small arenas in T and X can be explained by the fact that, when the time pressure is very low or the arena is small enough, it is optimal to perform only direct trajectories when certain, rather than hesitating. When time pressure increases (either by increasing T or X), the agent has now to start moving earlier and, when presented with contradictory evidence, either hesitating or waiting (thus increasing these signatures). At some point (which is different for waiting and hesitation for reasons explained above), the agent is so time pressured and/or the arena is so large, that she has to move directly and as soon as possible to the terminal points, therefore decreasing hesitation or post-start waiting.

3.4.9 Discussion

The major phenomenology observed in our experiments is that difficult trials elicit long decision trajectories (Figure 2.9 and 2.18), and that these trajectories appear to be time locked to the stimulus as seen in our analysis of Experiment 2 (Figure 2.21). Although this result suggests that the shape of the trajectories may contain information about an evolving decision variable during the trial, that analysis fails to incorporate how subjects' instantaneous movements are affected by their intended target, as well as their past

movements. To overcome this limitation we adopted a normative framework in which trajectories are considered optimal solutions to the problem of moving in a 2D environment towards one of two possible target locations. In each trial, a noisy time-varying sensory stream indicates which of the two locations will be rewarded, thus biasing the agent's trajectory. The optimization problem includes not only the reward associated with finding the correct location on each trial, but also costs associated with movement and time. We implemented this model by generalizing a similar approach used investigate the cost of time during evidence accumulation in a disembodied setting [Drugowitsch et al., 2012]. We tackle this reinforcement learning problem by numerically solving a Bellman equation in a state-space defined by time (within a trial), the agent's instantaneous belief about the correct choice, and its physical parameters. A major difference between the model and subject behavior in all of our experiments is that the model's actions are time locked to the stimulus, whereas the humans are free to move at any time, and successive moves are only constrained by the subject's own speed. We also implemented a model with the freedom to make several moves per time-step, but went forward with the simpler model without the extra refractory period parameter. This model's behavior was conceptually very similar to our first velocity constrained model attempting to simulate the arm based task. We observe regimes of cost parameters for which the optimal agent loosely tracks the stimulus (Figure 3.10). Our agent moves when her beliefs cross action thresholds (Figure 3.11), not unlike the bounds of the drift diffusion framework [Ratcliff, 1978], but in our framework new bounds on belief are encountered at each time and space in the task. The agent will move left (right) exactly one grid over to that side if her belief on the stimulus has crossed the left (right) decision bound, otherwise she will wait at her current spot. This model allows us to understand the actions of the agent in terms of their values to the agent. The agent is always comparing the long-term value of acting anytime later in the trial with the values of her immediate options available. This type of normative framework has been proposed before to explain down the middle reach trajectories [Haith et al., 2015]. In that work, the authors provide experimental and theoretical evidence that intermediate movements do not result from involuntary averaging between parallel motor plans, but instead are the result of an optimal calculation involving cognitive and motor constraints [Haith et al., 2015]. Their experiment is a go-before you know task, where reaction times are pushed to be very small

by placing explicit timing constraints similar to those we used in Experiment 2. Haith studied the effect of the distance between the targets on the trajectory decisions, and found that the incidence of intermediate movements depends on the target jump amplitude, with bigger amplitudes resulting in fewer intermediate amplitudes [Haith et al., 2015]. Another recent study similarly used a normative model of an embodied tasks in order to argue that intermediate movements result from purposeful planning which flexibly selects a single motor plan to maximize task success [Wong and Haith, 2017]. The effect of motor effort on changes of mind had been previously investigated by Burk [Burk et al., 2014]. The same group also recently used dynamic programming to model the change of mind and double change of mind phenomenology [Van Den Berg et al., 2016]. This last effort concludes rational decision makers adjust the bounds for the second decision of a double-decision task if there is a cost associated with the passage of time [Van Den Berg et al., 2016]. We too investigate a motor component to decision making using normative models, however our focus in on the geometric aspect is more similar to the work of [Haith et al., 2015]. However, as previously done by Drugowitsch and colleagues [Drugowitsch et al., 2012], we also investigate the effect of time systematically. Our focus on time and the sequential nature of behavior perhaps makes our study closest to the study of Vanden Berg [Van Den Berg et al., 2016]. With our model we described the effect of the time cost (Figure 3.18) and arena width (Figure 3.23) on the decision bounds at each time step of a multi-step decision, and ultimately the resulting behavior for a wide range of time-cost and arena sizes (Figure 3.24), while a controlled stochastic stimulus is presented through the entire duration of each trial. Our study combines elements from both of [Haith et al., 2015] and [Van Den Berg et al., 2016], providing a unified framework for the study of choice behavior in 2D embodied settings. Although a link to confidence has been made between the optimal model's beliefs and embodied actions, we do not explore confidence in our model. One of the most important aspects of the problem we studied is that it requires precise fine-tuning of the task conditions. This is evidenced by the difficulty in capturing hesitant behavior in experiment 1. The modeling work allows us to see that hesitation only occurs between the two asymptotic behaviors of the arenas (wide or long). The extreme shaped arenas don't give rise to hesitation and interesting trajectories, it is only in the intermediate regime where -for a given cost of time - the agent needs to start early, but still

has enough room to change if her early choice was premature, one observes patterned trajectories. In the model, hesitation is greatest for arena widths between 10 and 20. This can be observed in the top right panel of (Figure 3.22) and in the top left panel of (Figure 3.24). We think these figures explain the phenomenon we have been studying in our psychophysical work. In experiment 1, lack of time pressure for the time-scale of the stimulus we used resulted in little hesitation. In experiment 2, we increased the number of hesitant trajectories by imposing time pressure at start times. We use the simplified model to learn the ideal range to observe hesitant behavior, and used these observations to guide pilot experiments for this new task described in Chapter 4. Anecdotally, as in the model no large differences in the pattern of presses were seen in the arenas that had 7, 8, or 9 in human subjects. However, subjects didn't hesitate much for smaller arenas (e.g. 3 steps to each side) whereas the agent shows hesitation even for the smallest arenas. Model fitting was not an explicit goal of this analysis. Our initial goal was showing that such optimal consideration of space, in addition to temporal, constraints could generate in-between trajectories, in order to explore the range or parameters and inform the design of the final subject task. Model fitting might still be performed in the future either in the described model or in a non-dynamic programming version of the normative agent. Other more complex models were explored, but abandoned early as the idea of sequences of button presses was adopted. In the following chapter, we adapt our decision making task to be consistent with the rules of the environment of our embodied agent, using the arena sizes that maximize hesitant behavior in the model to guide the design of the subject task.

3.5 Acknowledgments

Chapter 3 was analyzed by me with the guidance of Alfonso Renart. We would like to thank Gautam Agarwal for providing the first normative model of our task, to Juan Castineiras for his derivation of an analytical equation for the Bayesian belief propagation driving reward expectation in the model, and Zach Mainen for discussions. A special thank you to Miguel Bengala for revising and re-plotting all the figures for the chapter under Juan Castineira's and my supervision.

Bibliography

[Burk et al., 2014] Burk, D., Ingram, J., Franklin, D., Shadlen, M., and Wolpert, D. (2014). Motor Effort Alters Changes of Mind in Sensorimotor Decision Making. *PloS one*, 9:e92681. (Cited on pages 98, 152, and 155.)

[DeepMind, 2019] DeepMind (2019). AlphaStar: Mastering the Real-Time Strategy Game StarCraft II. (Cited on page 59.)

[Drugowitsch et al., 2011] Drugowitsch, J., Moreno-Bote, R., and Pouget, A. (2011). Optimal decision bounds for probabilistic population codes and time varying evidence. *Nature Precedings*. (Cited on pages 9 and 59.)

[Drugowitsch et al., 2012] Drugowitsch, J., Moreno-Bote, R. N., Churchland, A. K., Shadlen, M. N., and Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. *Journal of Neuroscience*. (Cited on pages 6, 7, 9, 53, 58, 59, 60, 62, 65, 66, 74, 76, 78, 97, 98, 106, 153, and 154.)

[Google, 2015] Google (2015). AlphaGo | DeepMind. (Cited on page 59.)

[Haith et al., 2015] Haith, A. M., Huberdeau, D. M., and Krakauer, J. W. (2015). Hedging Your Bets: Intermediate Movements as Optimal Behavior in the Context of an Incomplete Decision. *PLoS Computational Biology*. (Cited on pages 97, 98, 145, 146, and 152.)

[Huang and Rao, 2013] Huang, Y. and Rao, R. P. (2013). Reward Optimization in the Primate Brain: A Probabilistic Model of Decision Making under Uncertainty. *PLoS ONE*. (Cited on pages 59, 60, and 61.)

[Moreno-Bote, 2010] Moreno-Bote, R. (2010). Decision confidence and uncertainty in diffusion models with partially correlated neuronal integrators. (Cited on pages 9, 59, 60, and 153.)

[Rao, 2010] Rao, R. (2010). Decision Making Under Uncertainty: A Neural Model Based on Partially Observable Markov Decision Processes. *Frontiers in Computational Neuroscience*, 4:146. (Cited on pages 59, 60, and 61.)

[Ratcliff, 1978] Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*. (Cited on pages 6 and 97.)

[Ratcliff et al., 2016] Ratcliff, R., Smith, P. L., Brown, S. D., and McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in cognitive sciences*, 20(4):260–281. (Cited on pages 6 and 59.)

[Silver et al., 2018] Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*. (Cited on page 59.)

[Silver et al., 2017] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., Van Den Driessche, G., Graepel, T., and Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*. (Cited on page 59.)

[Stone, 1960] Stone, M. (1960). Models for choice-reaction time. *Psychometrika*. (Cited on pages 6 and 59.)

[Sutton and Barto, 2018] Sutton, R. and Barto, A. (2018). Reinforcment Learning. *MIT Press*. (Cited on pages 58 and 153.)

[Van Den Berg et al., 2016] Van Den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N., and Wolpert, D. M. (2016). A common mechanism underlies changes of mind about decisions and confidence. *eLife*. (Cited on pages 98, 152, and 155.)

[Wald and Wolfowitz, 1948] Wald, A. and Wolfowitz, J. (1948). Optimum Character of the Sequential Probability Ratio Test. *Ann. Math. Statist.*, 19(3):326–339. (Cited on pages 6 and 59.)

[Wong and Haith, 2017] Wong, A. L. and Haith, A. M. (2017). Motor planning flexibly optimizes performance under uncertainty about task goals. *Nature Communications*. (Cited on pages 98 and 145.)

Chapter 4

A sequential decision making task for studying temporally extended actions

HIGHLIGHTS

- In this chapter we describe our final experiment, where the stimulus has been greatly simplified in order to reduce ambiguity regarding the subjective contribution of each stimulus sample, and human subjects now control the cursor with button presses in order to simplify the motor component of our task.
- As in our previous tasks and our model, average trajectory length increases significantly with increasing stimulus difficulty.
- Inter press interval decreases with increasing stimulus difficulty.
- Responses to repeated presentations of the same stimulus result in similar responses on different trials.
- Signatures of spatial and temporal dependence of actions can be observed in this task.

Contents

4.1 Abstract 106
4.2 Introduction 107
4.3 Methods 107
4.4 Results 112
4.5 Discussion 145
4.6 Acknowledgments 147

4.1 Abstract

Our previous theoretical work showed us that modifying the cost of accumulating evidence framework [Drugowitsch et al., 2012] to include a simple geometric cost for moving can result in agents that display hesitant decision trajectories. The discrete actions depending on nothing but an assumed cost per moving one unit step can be representative of human subjects pressing either of 2 buttons, and accordingly in this chapter we changed the input to our task from joystick movement to button presses. In addition to changing the motor input to the task, we also mimicked the cost of time of the model by presenting human subjects with clear visual feedback to inform subjects of their position in the arena, how much time remains, and the points lost due to having a constant cost for time passing. Subjects in our final experiment continue to display hesitant choice trajectories, and as in our previous experiments we find that harder stimuli are associated to longer choice trajectories. Analysis of the choice trajectories reveal stereotyped responses at the level of the time interval between presses, and also higher correlations between responses to repeats of the same stimulus when compared to responses to non-identical stimuli sampled from the same distribution. An effect of time passing on subject responses is present at the level of inter-press-intervals, which are slightly but significantly faster as the trial duration increases. Our models predict that the coupling between the stimulus and subject responses should become weaker in time, as agents eventually commit to the targets that they are closer to, regardless of what new evidence is presented to the agent.

4.2 Introduction

Although we were able to collect decision trajectories that were correlated to our stochastic stimulus in Experiment 2, accurate modeling of an optimal agent capable of displaying such trajectories became computationally prohibitive. Our modeling work allowed us to understand the role of time pressure in creating the environmental circumstances for which optimal agents hesitate. We realize it would be necessary to have a tighter control of the subjects' urgency to respond. To this end, we designed a follow up experiment where subject performance was explicitly linked to a point system, with rewards for correct responses and point reductions proportional to the passage of time. In this new task, subjects are told that the aim is to collect a point reward by moving a cursor in a grid to the corresponding target area. The reward is temporally discounted, and clearly indicated to the human subjects with feedback both online and during the inter-trial-interval. Subjects displace the cursor by one unit to the left (right) by pressing the left (right) button, which eliminated the need for modeling the derivatives of displacement and standardized responses across subjects. We also took this opportunity to simplify the stimulus in order to reduce subjective ambiguity regarding the contribution of each stimulus sample, which we changed to be binary (left or right) by eliminating small or zero ILD values. All of these changes allowed us to provide a data set in which subjects behave qualitatively similar to our optimal agents to Chapter 3. The use of repeated stimuli at random also allows us to quantify the degree to which subjects respond similarly to the same stimulus. This chapter provides a detailed quantification of subject behavior in Experiment 3.

4.3 Methods

4.3.1 Task

As in Chapter 3, the task is now to resolve the bias (left or right) of the stimulus before the time limit, and in view of temporally discounted rewards. The action space was modified such that subjects now move left or right by pressing one of two buttons at a time. However, unlike the model of Chapter 3, subjects are not limited to one button press per sample of stimulus, and

are free to move left or right as fast as they want, one press at a time. The task is self-initiated by pressing (and releasing) the left and right button simultaneously, which triggers a random delay of 500 ms - 1000 ms, after which the trial and the stimulus start. The trial ends when subjects reach one of the two distally separated targets (or walls), or when the time limit is reached. A cursor is displayed and automatically moved upwards towards an upper bound at constant speed, signaling the time and points remaining. The cursor itself fills up with a different color to convey the same information about time in the task.

4.3.2 New Stimulus

In this experiment, we used the stimulus described in the modeling chapter (3.3.4.1). Initially, 10,000 stimuli were created by setting p = 0.6 for generating a right favoring sample, and left stimuli were created by mirroring the right stimuli. However, since sampling with a probability of 0.6 does not necessarily result in right favoring stimulus sequences over the entire duration of the stimulus(28 samples or 7 seconds), 6 different difficulty conditions were created by selecting stimuli with specific proportions of left to right samples over the 7 seconds of stimulus. A total of 9 unique leftward stimuli were kept per session (3 per each of the 3 different left difficulties), and mirrored for the left. Each session used a different seed for the first 5 sessions over one week, and the seeds were repeated on the second week. See (4.3.5) Subjects and Sampling Procedure for more details.

4.3.3 Pilot 2AFC Fixed Duration Experiment

In order to confirm that the discrimination of this stochastic stimulus is an integration task, we ran a series of standard 2AFC fixed duration experiments, progressively allowing subjects to experience the stimulus for shorter durations. We expect the sessions with longer durations to have better performance, and this is indeed what we saw.

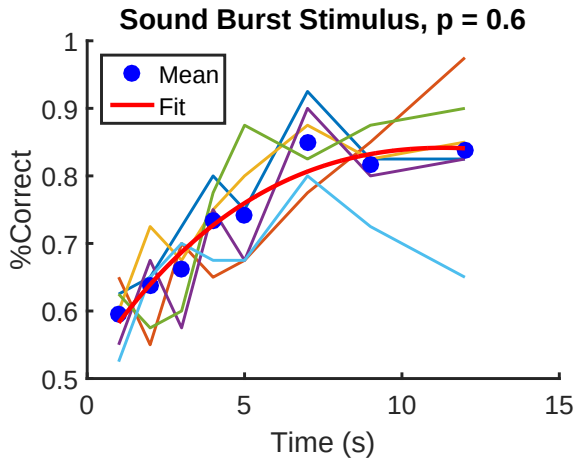


Figure 4.1: Fixed duration 2AFC experiments. We tested 6 naive subjects (all different from the main experiment subjects), with only the hard conditions, $p = 0.6$ (Right Stimulus) and $p = 0.4$ (Left Stimulus), at different fixed durations of 1, 2, 3, 4, 5, 7, 9 , and 12 seconds in separate session blocks of 40 trials per duration with a break in between. As expected, as subjects were allowed to experience shorter stimuli their accuracy decreased

.

4.3.4 Temporally Discounted Reward

Aside from the changes in the stimulus, we also thought it necessary to set an explicit cost for waiting in the psychophysics task to match the type of costs imposed on the model. We put a constant cost per unit of time, which results in an accumulated cost for waiting that grows linearly in time. We set this quantity such that the accumulated time cost would match the total rewards by the end of the task. Theoretically, subtracting this cost from the monotonically growing value of the expected reward in time (as the subjects become more certain), this should result in an expected value for completing the task that reaches its maximum value sometime in the middle of the task. A schematic of this reward system is shown below (Figure 4.2).

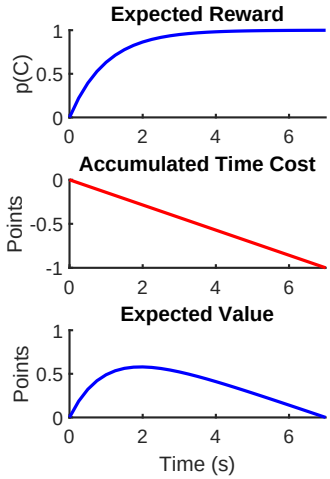


Figure 4.2: This is a schematic of the expected reward, which grows in time as subjects accumulate more evidence and increase their probability of answering correctly (top), a constant time cost which results in a linearly accumulating point reduction (middle), and the addition of these two quantities resulting in the expected value for completing the task as a function of time (bottom).

4.3.5 Subjects and Experimental Procedure

A total of 4 subjects completed the entire experiment, following 1 training session. In order to assess the reliability of the responses to a specific stimilus, we repeated stimuli in a random way to ensure subjects didn't realized they were listening to the same stimulus. Each week subjects were asked to complete 5 sessions (one per day) of 72 trials each. Each session consisted of 18 unique stimuli (9 unique stimuli to the right were created and then flipped to make the left stimuli), and each was repeated 4 times per session in a random order. Subjects completed 10 sessions across 2 weeks, and the exact stimuli and order or week 1 were repeated on week 2, by setting a unique random seed for each day of the week. Subjects were explicitly asked at the end of the entire experiment if they noticed the repeats, and none reported having noticed. In total, each particular stimulus was experienced 8 times per subject. This resulted in 90 unique sequences (45 to the right mirrored for the left conditions, with 8 repetitions per sequence), totalling 720 trials per subject. Sample responses for 4 subjects and all repeats of

some sample stimuli are shown in (Figure 4.3).

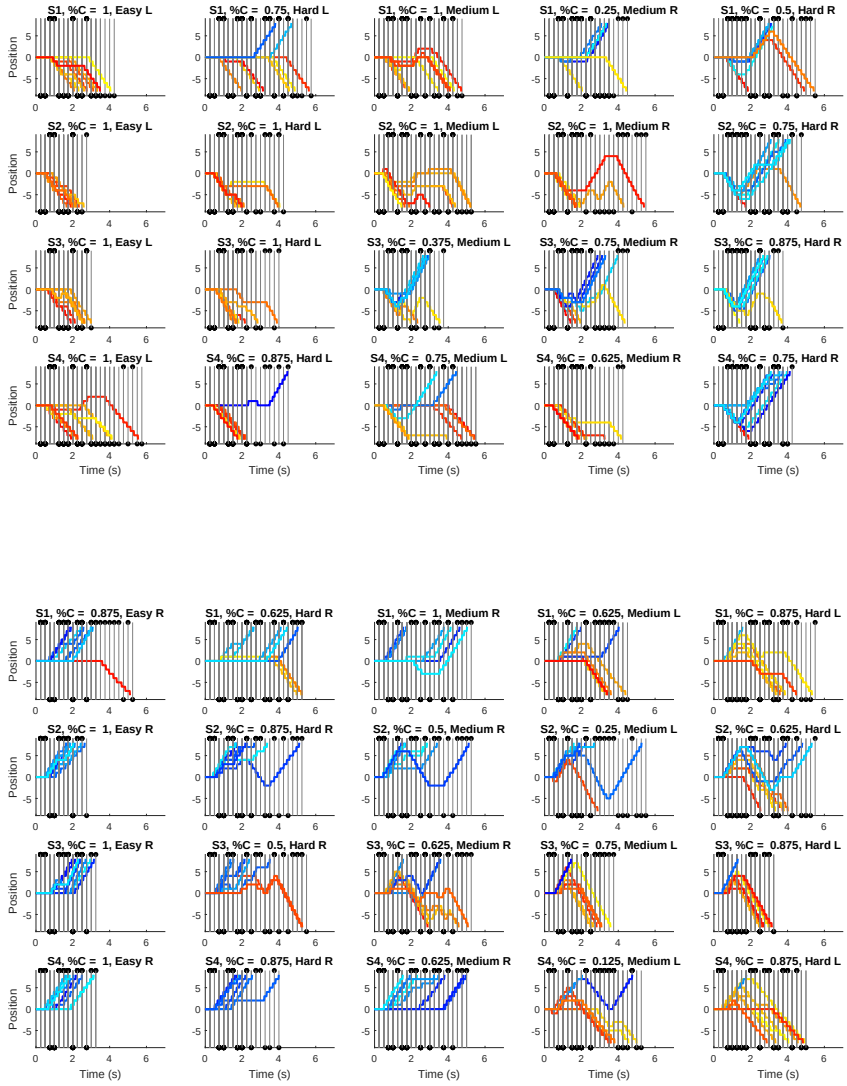


Figure 4.3: Stimulus Repeats. Top panel: Each column shows a unique leftward stimulus and 8 responses from each subject. Each row shows a different subject. Bottom panel: Mirrored rightward stimuli. Black dots on each plot denote the stimulus. Blue (red) colored trajectories end on the right (left).

4.4 Results

4.4.1 Task Stability

Percentage correct, Scored Points, and Reaction times do not trend across sessions, suggesting performance is stable across sessions for all 4 subjects.

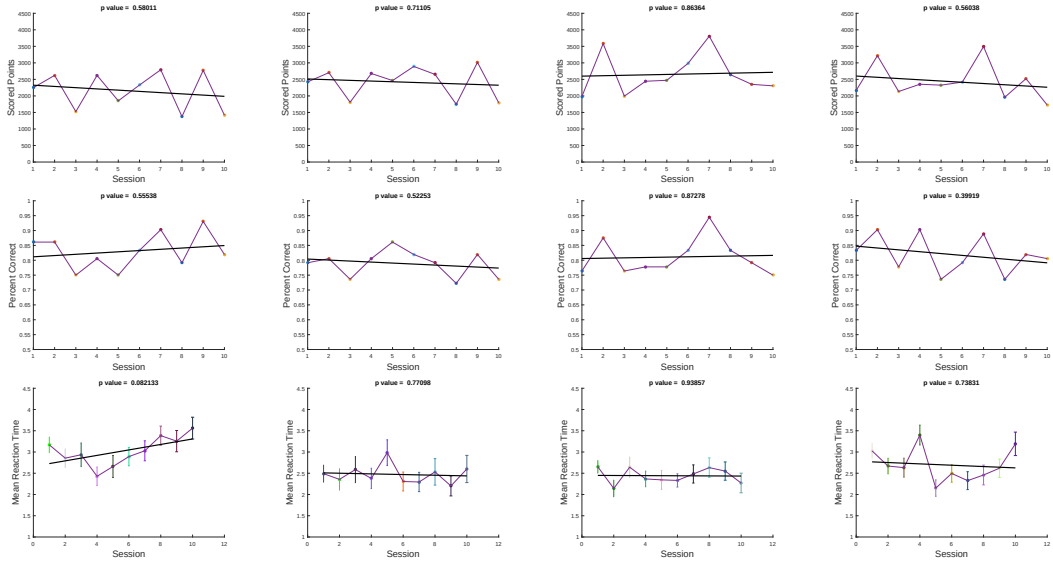


Figure 4.4: Scored Points (top row), Percentage correct (middle row), and Reaction times (bottom row) are plotted against session number for all subjects in order to investigate whether any linear trends exist, which would suggest learning is still ongoing. No such trends can be detected at a significance of $\alpha = 0.5$, suggesting performance is stable across sessions. Error bars show standard deviations.

4.4.2 Accuracy

Psychometric curves for 4 subjects are shown below, defining conditions according to the long-term mean of the stimulus.

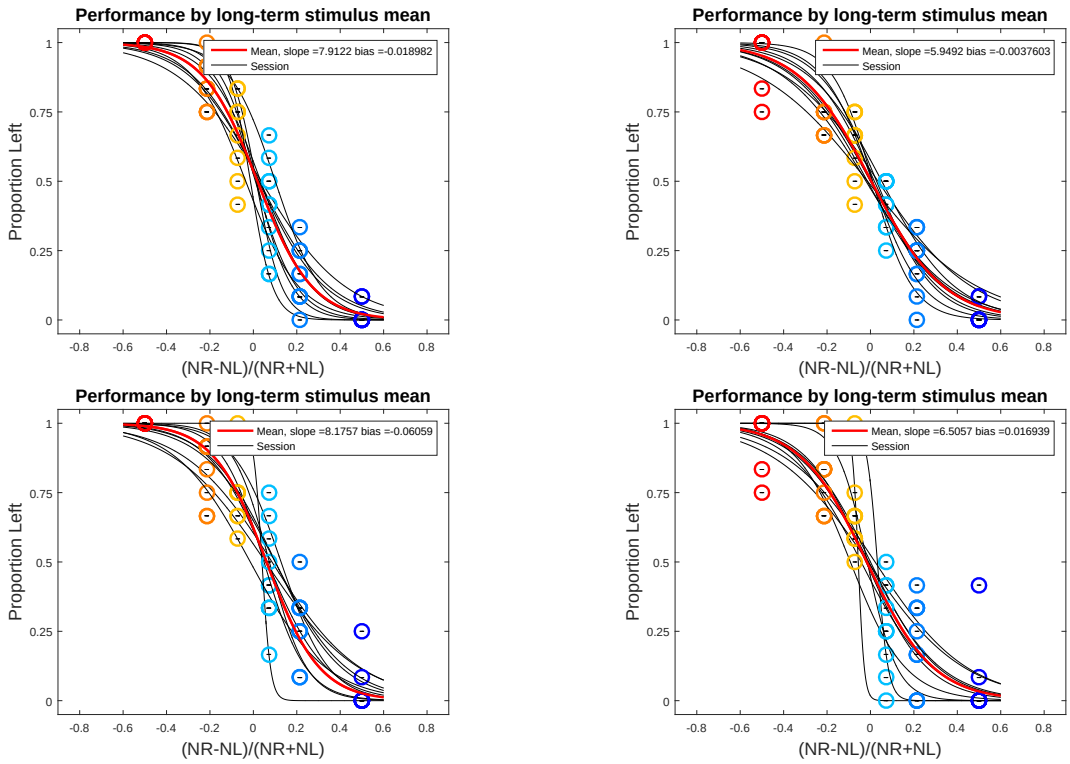


Figure 4.5: Psychometric curves for 4 subjects evaluating performance according to the long-term mean of the stimulus. Thin black lines show performance for each of the 10 sessions of 72 trials, thick red lines show average psychometric over all 720 trials for each subject. Red to blue graded circles show session performance for 6 difficulty conditions.

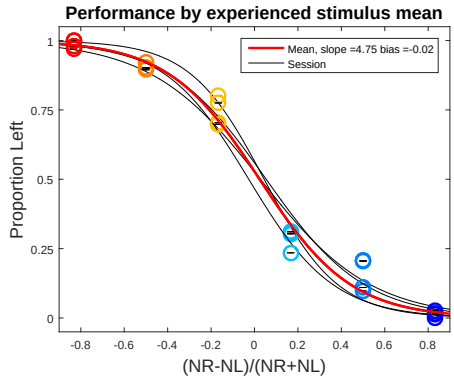


Figure 4.8: Average psychometric curve for 4 subjects evaluating performance according to the mean of the experienced the stimulus. Thin black lines show performance for each subject over 720 trials, thick red lines shows the mean for all subjects. Red to blue graded circles show session performance for 6 difficulty conditions.

4.4.3 Reaction Times

Reaction times for 4 subjects increase as a function of the difficulty of the experienced stimulus.

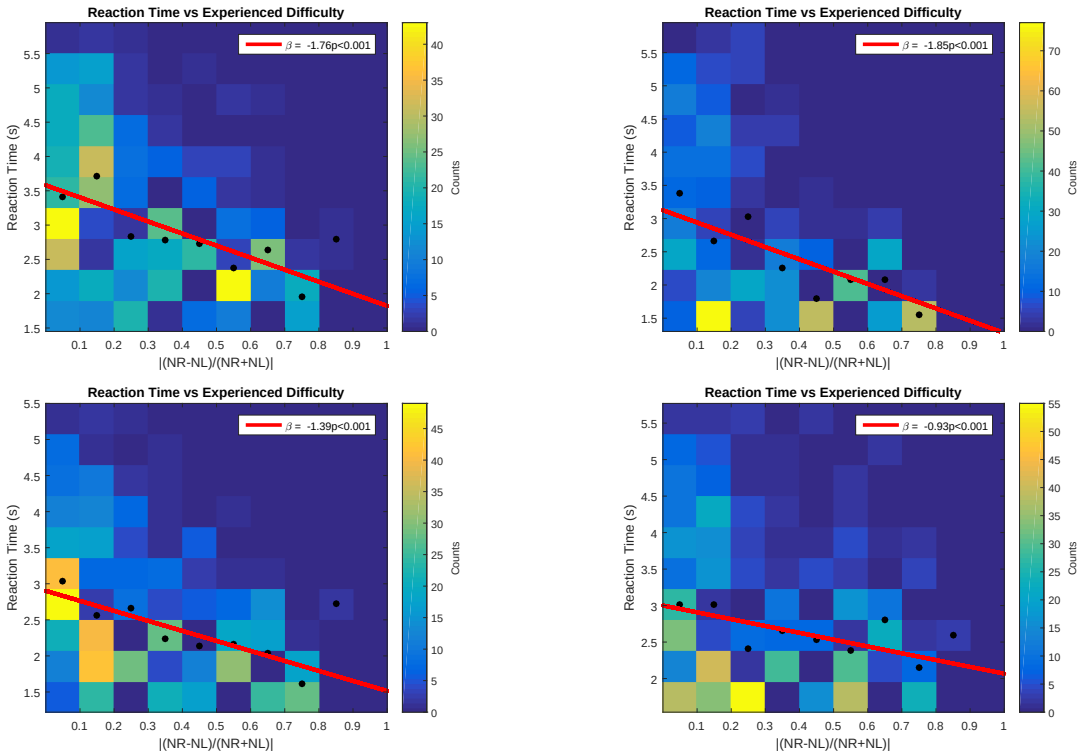


Figure 4.9: Reaction times for 4 subjects decrease as the mean of the experienced stimulus increases (task becomes easier) (3.23). Black dots show mean reaction times for each difficulty bin, red line shows the linear fit of RT onto stimulus difficulty, and the color bar the counts of RT for each bin.

In order to compute the mean reaction time vs stimulus difficulty we collapse all subject reaction time data without normalization. The choice of not normalizing is due to the similarity in shape and magnitude of the reaction time distributions across subjects,

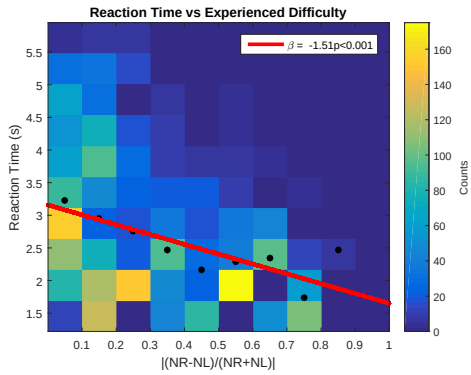


Figure 4.10: Reaction times decrease as the mean of the experienced stimulus increases (task becomes easier) (3.23). 2280 trials collapsed across all 4 subjects. Black dots show mean reaction times for each difficulty bin, red line shows the linear fit of RT onto stimulus difficulty, and the color bar the counts of RT for each bin.

4.4.4 Trajectory Analysis

4.4.4.1 Trajectory Length

The main feature observed throughout our experiments is that the decision trajectory length increases as the stimulus becomes hard, and the model predicts that we should continue to observe this behavior with discrete button press input to the task - for a range of time costs. In our new task, the number of presses increases as a function of the difficulty of the experienced stimulus for all 4 subjects (Figure 4.11). We used 7 seconds and an arena of width x17, parameters that were in the region of maximum Hesitation for the model and worked well in pilot experiments.

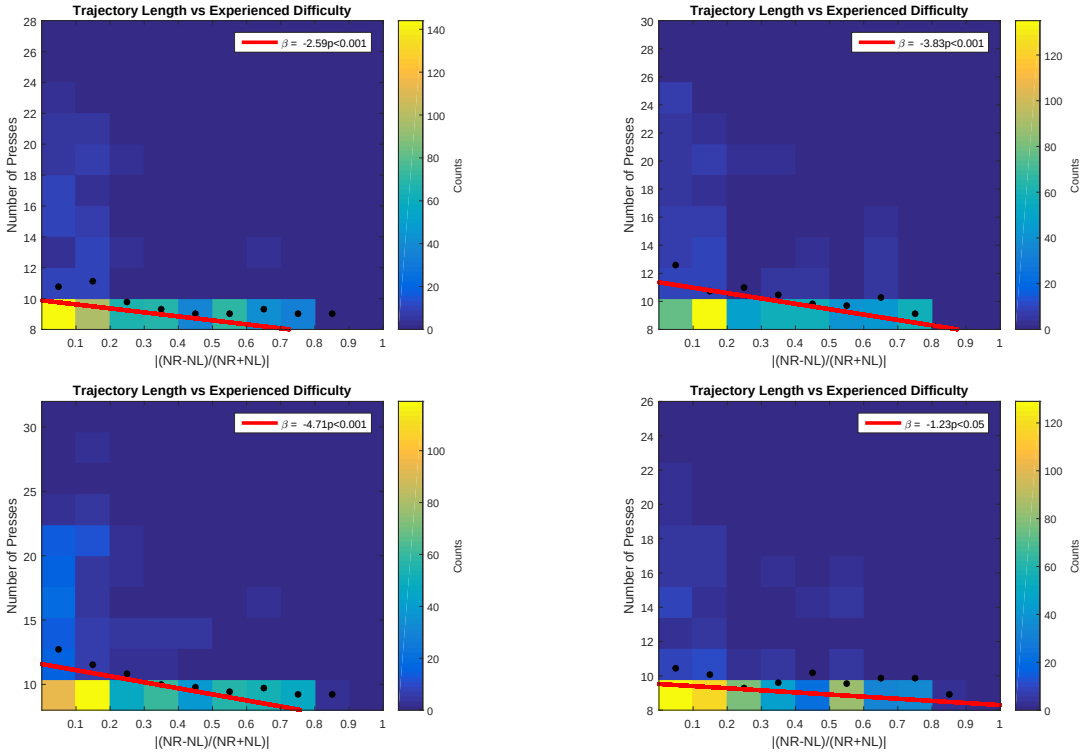


Figure 4.11: The number of presses across all 4 subjects increases with increasing difficulty of the experienced stimulus (3.23 → 0). Black dots show mean Number of Presses for each difficulty bin, red line shows the linear fit of Number of Presses onto stimulus difficulty, and the color bar the counts of Number of Presses for each bin.

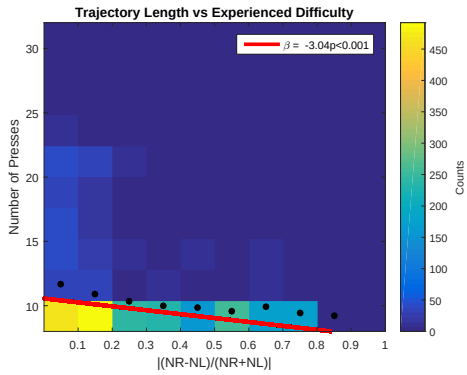


Figure 4.12: Number of presses increase with increasing difficulty of the experienced stimulus (3.23 → 0). 2280 trials collapsed across all 4 subjects. Black dots show mean Number of Presses for each difficulty bin, red line shows the linear fit of Number of Presses onto stimulus difficulty, and the color bar the counts of Number of Presses for each bin.

4.4.4.2 Inter-press Interval Reliability

We observed that across subjects, the dristribution of inter-press-intervals (IPI) was very similar, and peak IPIs were almost overlaping.

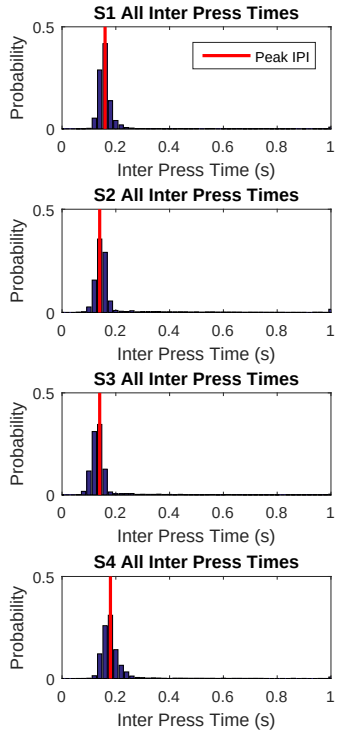


Figure 4.13: Each row shows one subject's distributions of inter-press intervals (IPI) for all responses. Red line shows the peak IPI.

Furthermore, we observed this pattern for all subjects and for all different difficulty conditions, inter-press interval distributions are concentrated around a peak of <200 ms. We initially thought this observation suggested that subjects were always in one of three states: waiting, press and continue pressing left, or press and continue pressing right.

However, upon further inspection, we noticed that for harder trials, longer IPIs were present shifting the 95 percentile of the distribution towards higher values. This suggests a 4th state, uncomitted pressing.

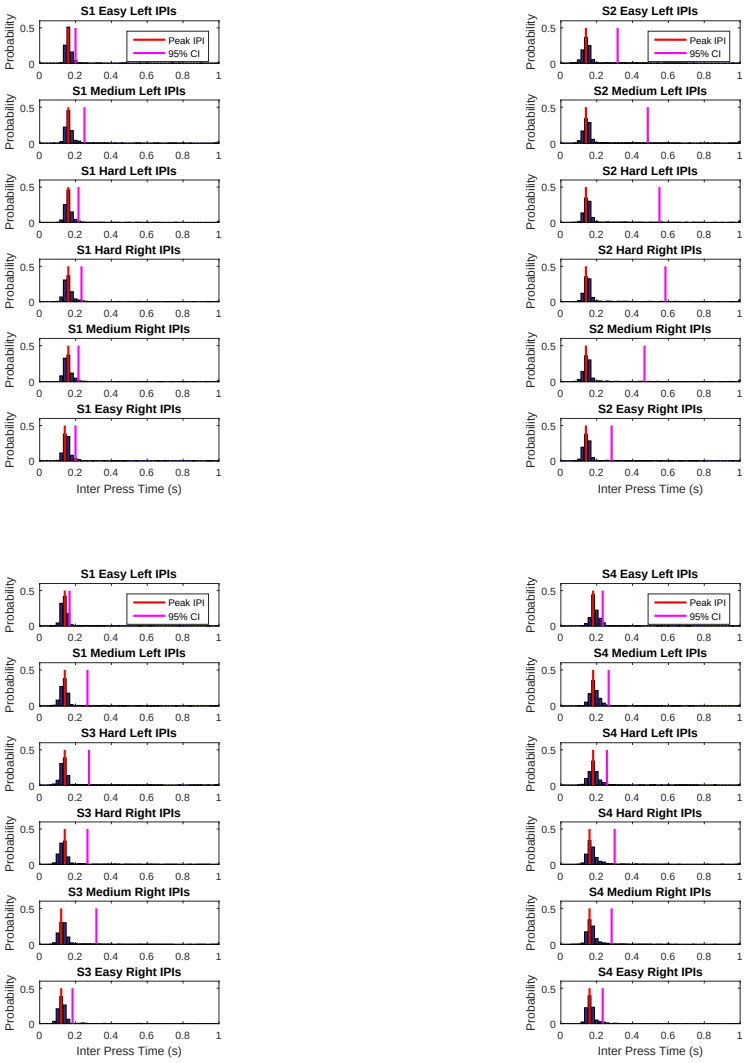


Figure 4.14: Inter-press-interval (IPI) distributions for all subjects (each column) across 6 different difficulty conditions (each row). Red line marks the peak, magenta line marks the 95th percentile.

Comparing the IPI to stimulus difficulty allows us to see that in fact, IPIs decrease with decreasing trial difficulty.

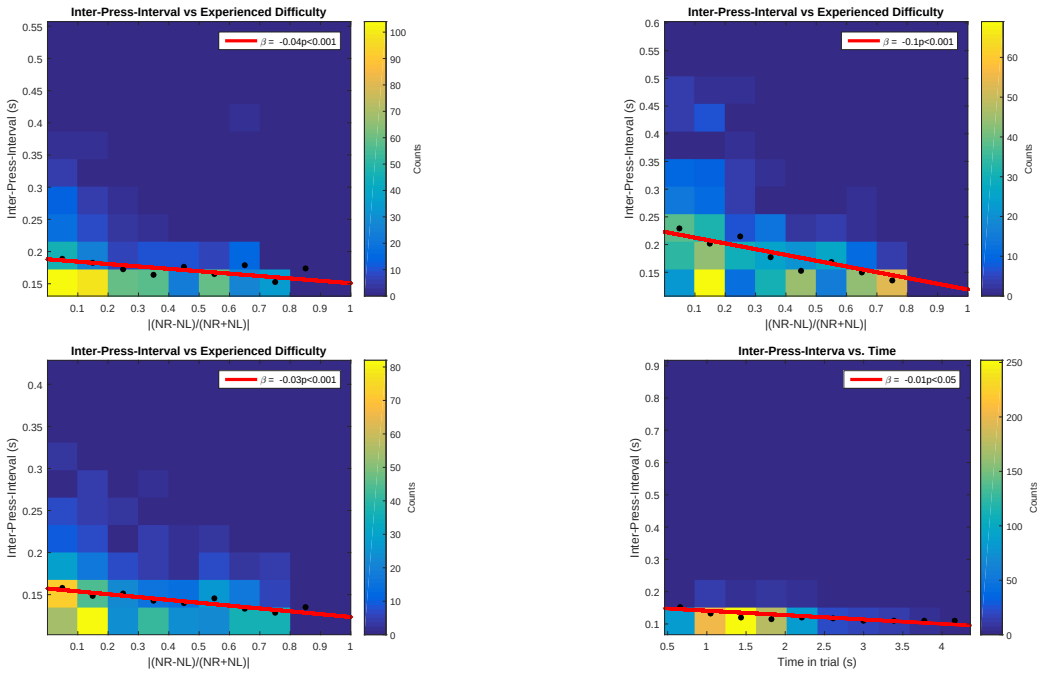


Figure 4.15: Across all 4 subjects inter-press-interval (IPI) decreases with decreasing difficulty of the experienced stimulus (3.23 → 1). Black dots show mean IPI for each difficulty bin, red line shows the linear fit of IPI onto stimulus difficulty, and the color bar the counts of IPI for each bin.

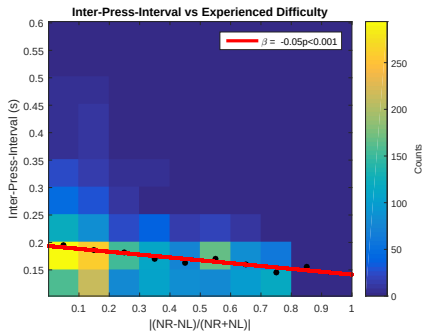


Figure 4.16: Inter-press-interval decreases with decreasing difficulty of the experienced stimulus (3.23 → 1). 2280 trials collapsed across all 4 subjects. Black dots show mean IPI for each difficulty bin, red line shows the linear fit of IPI onto stimulus difficulty, and the color bar the counts of IPI for each bin.

No differences in the above pattern were observed when analyzing correct and incorrect trials separately (not shown).

In order to assess whether in addition to the action states 1) pressing at or near peak IPI to the left 2) pressing at or near peak IPI to the right, and 3) waiting, there is a fourth state, we repeat this analysis removing any IPI values greater than 0.5 seconds, arbitrarily chosen as a threshold for waiting since it is more than 2x longer than the peak IPI but not so long that the analysis is a pointless repeat of (Figure 4.16). We find that the regression coefficient relating IPI to difficulty is virtually unchanged by the removal of waiting. We thus conclude that subjects have a fourth action state, hesitant pressing depending on the stimulus mean.

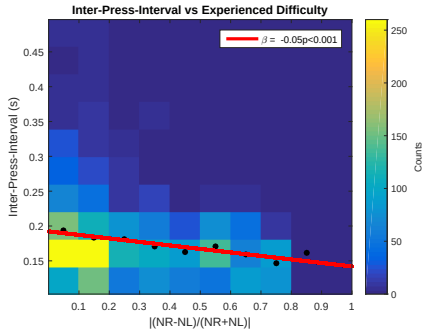


Figure 4.17: Even after the removal of any IPIs > 0.5 seconds, inter-press-interval decreases with decreasing difficulty of the experienced stimulus (3.23 → 1). 2280 trials collapsed across all 4 subjects. Black dots show mean IPI for each difficulty bin, red line shows the linear fit of IPI onto stimulus difficulty, and the color bar the counts of IPI for each bin.

In light of the modeling result that agents become more committed as time passes, we wondered whether the pattern of IPIs could show a time dependence (note that the model does not have varying IPIs). In order to assess the time dependence of IPIs, we analyze them as a function of the time when they occurred in the trial. Regression analysis shows a small tendency for longer trials to have slightly shorter IPIs. The effect is small, but significant for 3 of the 4 subjects (Figure 4.19) This is the first hint of a time dependence in the action space of the subjects - beyond reaction time itself - in this task. This is meaningful because our normative model becomes more committed as time progresses in a trial, and we expect some signature(s) of this in subject behavior.

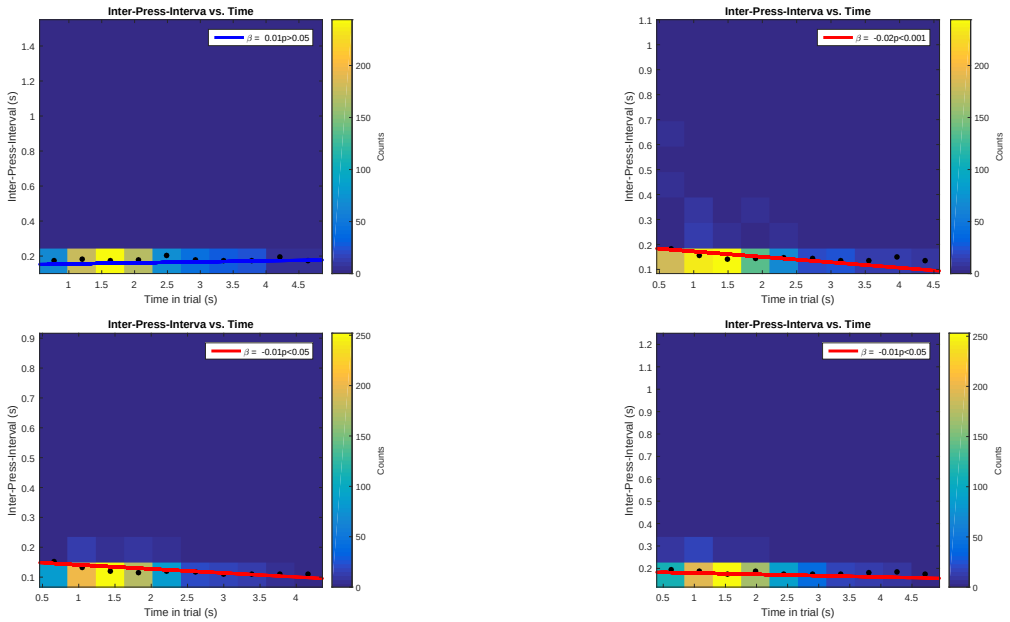


Figure 4.18: For 3 of 4 subjects, inter-press-interval decreases with increasing duration in the trial. Black dots show mean IPI for each difficulty bin, red line shows the linear fit of IPI onto stimulus difficulty, and the color bar the counts of IPI for each bin.

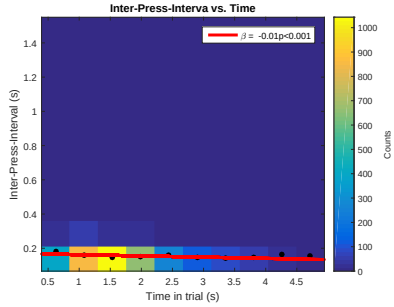


Figure 4.19: Inter-press-interval decreases with increasing duration in the trials. The effect is very small but significant for 2880 trials across 4 subjects. Black dots show mean IPI for each difficulty bin, red line shows the linear fit of IPI onto stimulus difficulty, and the color bar the counts of IPI for each bin.

4.4.4.3 Trajectory Reliability

In order to abstract the apparent reliability of responses to repeats of the same stimulus (Figure 4.3) into a summary statistic, we compute pairwise correlations between trajectories elicited by stimulus repeats vs trajectories elicited by stimuli sampled from the same difficulty condition. For each of 90 unique stimuli, we produce such a matrix, sorting it such that the first 8 elements correspond to the repeats of the same stimulus, and elements 8-16 show the randomly chosen stimuli sampled from the exact same difficulty condition, p(r). A sample matrix can be seen in (Figure 4.20), and 2 more in (Figure 4.21). For statistical analysis of these matrices see (Figure 4.22).

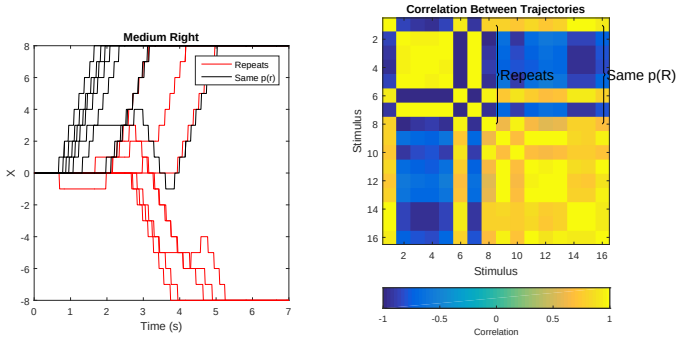


Figure 4.20: Left panel shows 8 responses to repeats of the same stimulus in red, and 8 responses randomly chosen from the 112 responses to the other 14 unique stimuli with the same difficulty condition, p(R), in black (note that these other 112 responses contain 8 repeats of each of the other 14 stimuli with the same condition). Right panel shows the pairwise correlations of the responses shown on the corresponding left panel. Colorbar is the correlation. The matrix is sorted such that the first 8 elements correspond to responses of repeats of the same stimulus, and elements 9-16 show responses to randomly chosen stimuli sampled from the exact same difficulty condition, p(r).

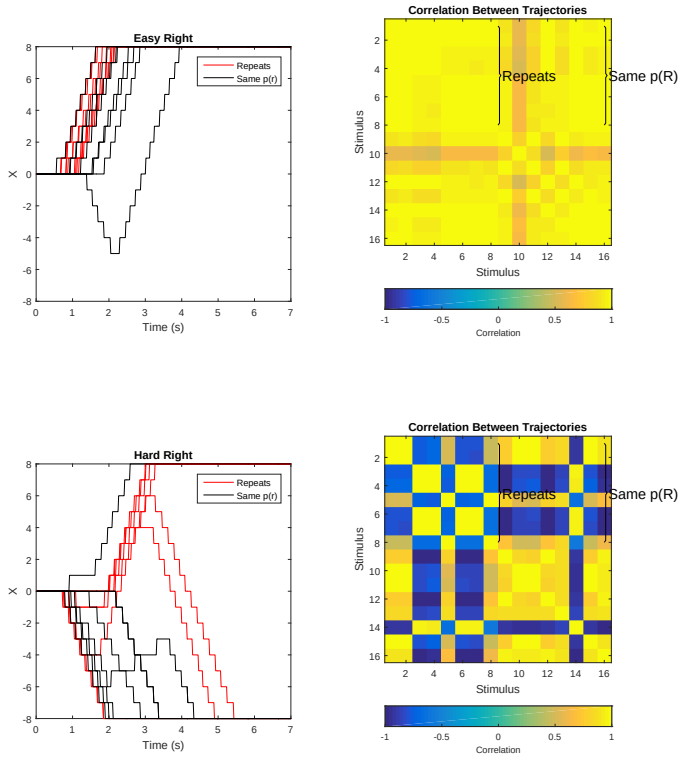


Figure 4.21: Each left panel shows 8 responses to repeats of the same stimulus in red, and 8 responses randomly chosen from the 112 responses to the other 14 unique stimuli with the same difficulty condition, p(R), in black (note that these other 112 responses contain 8 repeats of each of the other 14 stimuli with the same condition). Each right panel shows the pairwise correlations of the responses shown on the corresponding left panel. Colorbar is the correlation. The matrices are sorted such that the first 8 elements correspond to responses of repeats of the same stimulus, and elements 8-16 show responses to randomly chosen stimuli sampled from the exact same difficulty condition, p(r).

The correlations observed for repeats vs same p observed in (Figure 4.20) are analyzed in detail and as a function of the difficulty of the stimulus below (Figure 4.22).

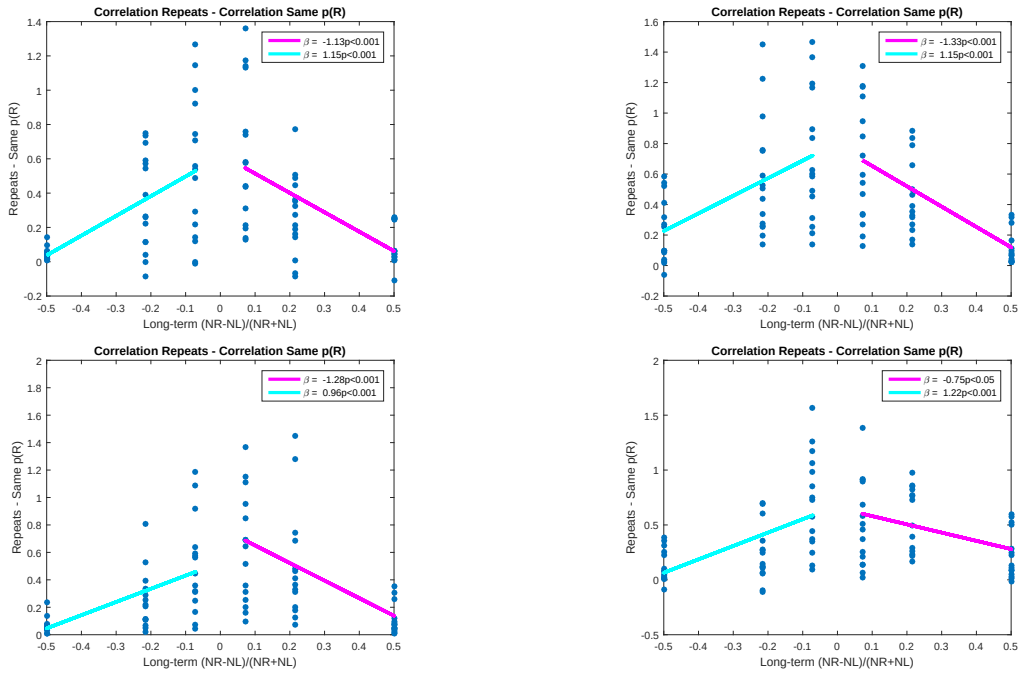


Figure 4.22: For all 4 subjects, difference between the mean of the same-stimulus correlations sub-matrix and the mean of the same-condition correlations sub-matrix depends on the difficulty of the stimulus. For each unique stimulus matrix of correlations A, we create 2 sub-matrices, one of the correlations between the responses to repeats, Repeats = A(1:8, 18), and one of the correlations between each unique repeat response and the 8 same p(R) condition responses, Same Conditions = A(1:8, 9:16). We eliminate identity and duplicate entries from the Repeats sub-matrix by further sub-sampling only the upper triangular matrix. Finally, we subtract the mean of the Same Conditions sub-matrix from the mean of the Repeats sub-matrix and find that this difference depends on the difficulty of the stimulus. Cyan (magenta) line shows the linear regression of the difference between the mean repeats correlations and the mean same conditions correlation onto the stimulus strength for leftward (rightward) trials.

As an alternative analysis, we average all 90 matrices for each subject. A clear difference between the correlations between repeats vs same condition stimuli can be seen (Figure 4.23).

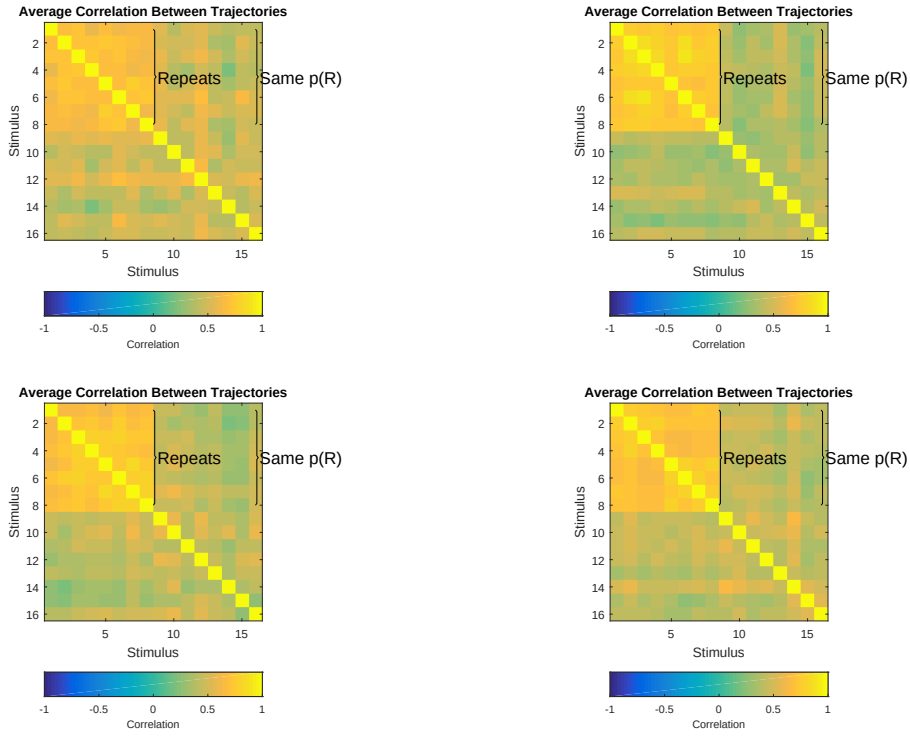


Figure 4.23: For each subject, an average correlations matrix is computed by averaging all 90 pairwise correlations matrices comparing responses to stimulus repeats to responses of stimuli sampled from the same condition. Matrices are sorted such that the first 8 elements correspond to responses of repeats of the same stimulus, and elements 9-16 show responses to randomly chosen stimuli sampled from the exact same difficulty condition, p(r).

Statistical comparisons at the level of the average matrices (Figure 4.23) is performed via Monte-Carlo Simulations. We apply a similar sub-set procedure to the average matrices as that followed in (Figure 4.22). The goal is to compare the mean of a sub-matrix of average correlations between repeated stimuli, T-stat, to the mean of a sub-matrix of average correlations between each unique repeat and 8 stimuli sampled from the same condition. We repeat this procedure 1000 times in order to generate a distribution of surrogate mean average correlations to same-condition stimuli, which can be used to compute the probability of observing T-stat by chance. Each subject's average matrix of correlations (Figure 4.23), B, is subset to cre-

ate 2 sub-matrices, one of the average correlations between the responses to repeats, Repeats = B(1:8, 18), and one of the average correlations of all the unique repeat response and the 8 same p(R) condition responses, Same Conditions = B(1:8, 9:16). We eliminate identity and duplicate entries from the Repeats sub-matrix by further sub-sampling only the upper triangular matrix. We statistically compare the value of the mean of Repeats, our test statistic, to that of 1000 surrogate values of the mean of Same Conditions and observe a significant difference for all subjects ($p<0.001$ bootstrapping test).

These analysis continue to give significant results when collapsing the data across all 4 subjects.

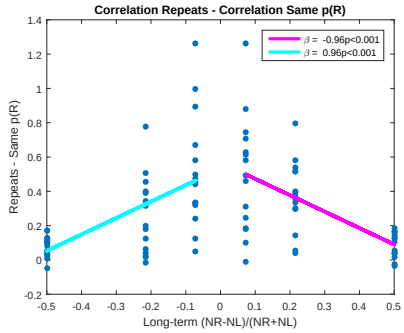


Figure 4.24: The difference between the mean of the same-stimulus correlations sub-matrix and the mean of the same-condition correlations sub-matrix depends on the difficulty of the stimulus, data collapsed across 4 subjects. Cyan (magenta) line shows the linear regression of the difference between the mean repeats correlations and the mean same conditions correlation onto the stimulus strength for leftward (rightward) trials.

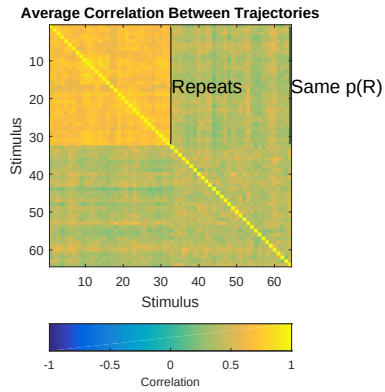


Figure 4.25: Average correlations matrix is computed by averaging all 90 pairwise correlations matrices comparing responses to stimulus repeats to responses of stimuli sampled from the same condition, collapsed across all 4 subjects. Matrices are sorted such that the first 32 elements correspond to responses of repeats of the same stimulus, and elements 33-64 show responses to randomly chosen stimuli sampled from the exact same difficulty condition, p(r). The difference between the mean of the same-stimulus correlations sub-matrix and the mean of the same-condition correlations sub-matrix is significant for all subjects ($p<0.001$ bootstrapping test).

4.4.4.4 Commitment: Space-time dependence of Actions

For our final set of analysis, we try to show the dependence of the action space of the subjects on their position in the arena and time in the task. From our pilot disembodied experiment (Figure 4.1), we know that in a standard 2AFC configuration the subjects integrate our stimulus in time. In the embodied task, our model from Chapter 3 suggests that as the bounds on evidence narrow in time and collapse in the direction of the agent's movement, agents become less tuned to the stimulus and their probabilities of waiting become smaller while the probabilities of pressing increase towards the side the agent of the arena the agent is in. Thus, we try to empirically calculate p(L,W,R|E), the probabilities of making a left movement, waiting, or making a right movement given the accumulated evidence the subjects have heard at 5 different position intervals in the arena and 3 different time intervals. Evidence is defined as in (3.3.4.3) over a window or time spanning N samples in the past, but excluding the last M samples, since in humans

we expect a delay in processing the stimuli to guide actions. Start times for the task are around 0.5-1 seconds, suggesting M is between 2 and 4. N<5 often did not produce complete histograms, and N between 5 and 8 tend to produce similar results. For some N and M parameter combinations, occasionally the evidence 3.3.4.3 does not span all the bins in the axis (see white space in the center bottom row histogram for subject 3 (Figure 4.28) and subject 4 (Figure 4.29).

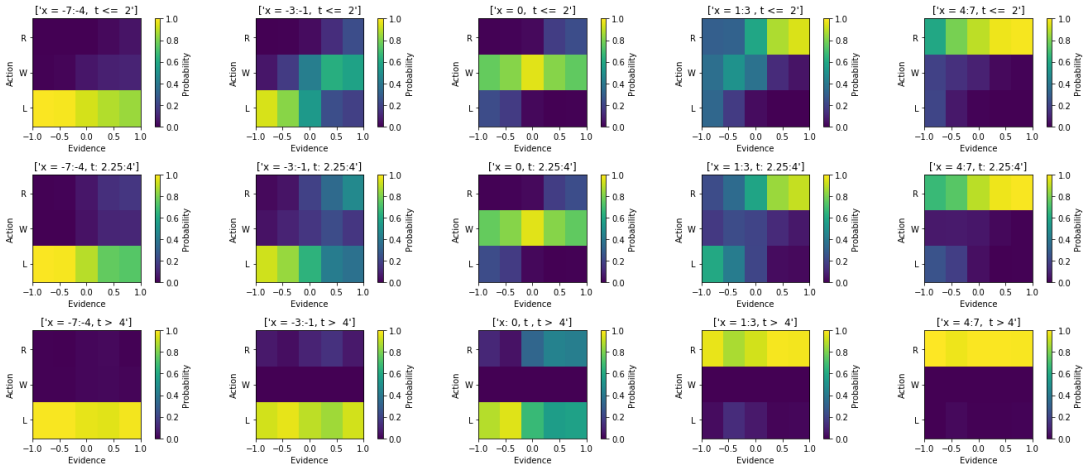


Figure 4.26: P(L,W,R|E) at different locations and time intervals in the arena for subject 1. For each row, the leftmost (rightmost) plots represent the leftmost (rightmost) spaces in the arena, and the center plot is the center of the arena. Time is split from 0 to 2 seconds (top row), 2 to 4 seconds (middle row), and 4 to 7 seconds (bottom row). At each time step, we count the number of times the subject waited, moved left, or moved right within the next 0.25 seconds, given how much evidence the subject has experienced. Evidence is defined as 3.3.4.3 over the last N samples, excluding the latest M to account for the delay between experiencing a stimulus and making a deliberate press. Results are shown for N = 7 and M = 3.

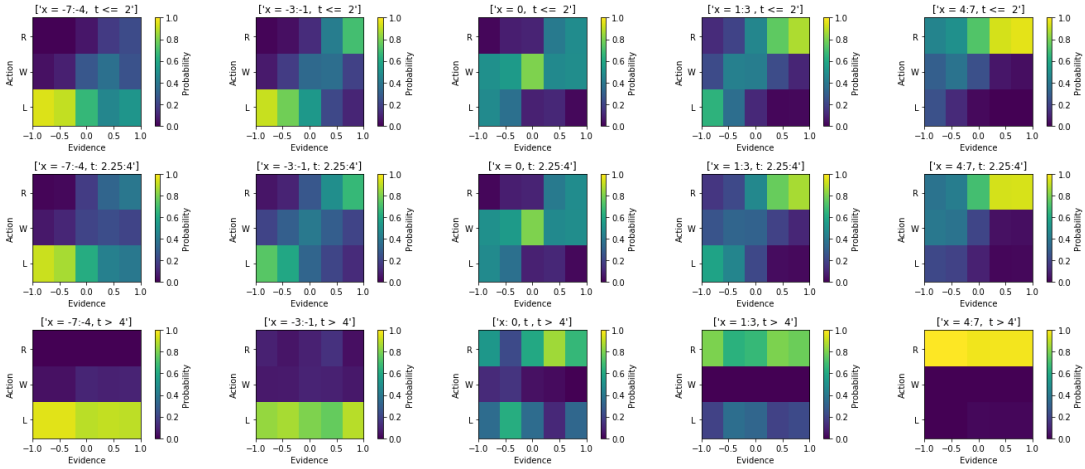


Figure 4.27: P(L,W,R|E) at different locations and time intervals in the arena for subject 2. For each row, the leftmost (rightmost) plots represent the leftmost (rightmost) spaces in the arena, and the center plot is the center of the arena. Time is split from 0 to 2 seconds (top row), 2 to 4 seconds (middle row), and 4 to 7 seconds (bottom row). At each time step, we count the number of times the subject waited, moved left, or moved right within the next 0.25 seconds, given how much evidence the subject has experienced. Evidence is defined as 3.3.4.3 over the last N samples, excluding the latest M to account for the delay between experiencing a stimulus and making a deliberate press. Results are shown for N = 7 and M = 3.

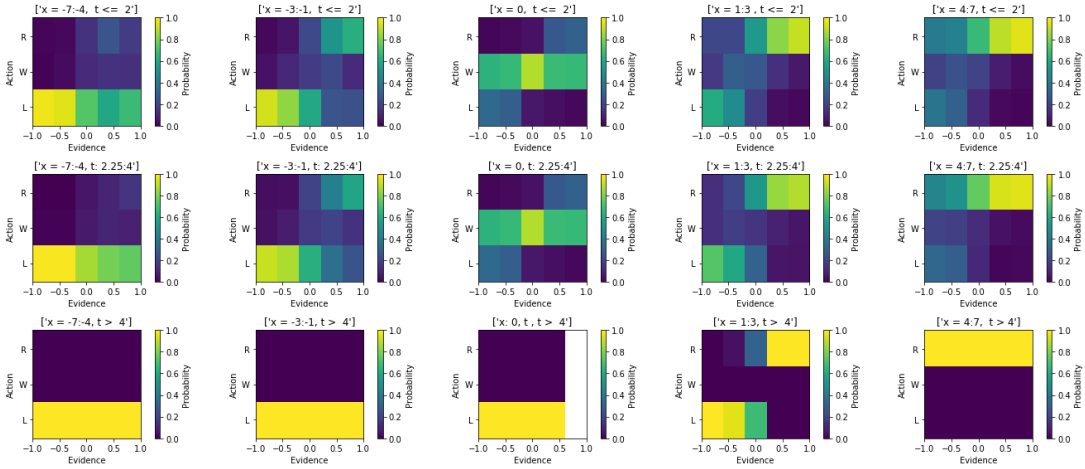


Figure 4.28: P(L,W,R|E) at different locations and time intervals in the arena for subject 3. For each row, the leftmost (rightmost) plots represent the leftmost (rightmost) spaces in the arena, and the center plot is the center of the arena. Time is split from 0 to 2 seconds (top row), 2 to 4 seconds (middle row), and 4 to 7 seconds (bottom row). At each time step, we count the number of times the subject waited, moved left, or moved right within the next 0.25 seconds, given how much evidence the subject has experienced. Evidence is defined as 3.3.4.3 over the last N samples, excluding the latest M to account for the delay between experiencing a stimulus and making a deliberate press. Results are shown for N = 7 and M = 3.

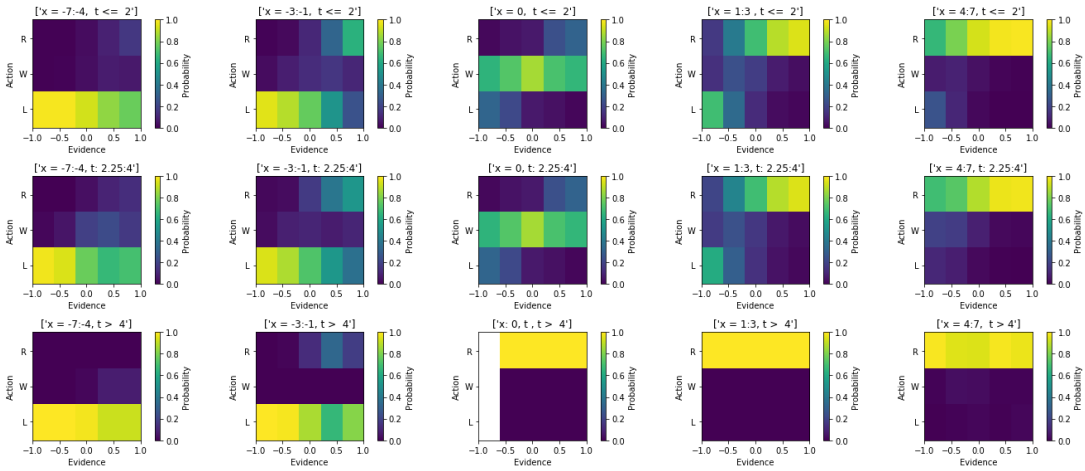


Figure 4.29: P(L,W,R|E) at different locations and time intervals in the arena for subject 4. For each row, the leftmost (rightmost) plots represent the leftmost (rightmost) spaces in the arena, and the center plot is the center of the arena. Time is split from 0 to 2 seconds (top row), 2 to 4 seconds (middle row), and 4 to 7 seconds (bottom row). At each time step, we count the number of times the subject waited, moved left, or moved right within the next 0.25 seconds, given how much evidence the subject has experienced. Evidence is defined as 3.3.4.3 over the last N samples, excluding the latest M to account for the delay between experiencing a stimulus and making a deliberate press. Results are shown for N = 7 and M = 3.

We observe that P(L,W,R|E) changes as subjects move in the arena, shifting from waiting to pressing towards the side of the arena the subject is in. Focusing on one row of panels and moving across the columns changes the maximum of P(L,W,R|E), shifting it towards the side that the subject is in. This pattern gets stronger as time advances in the task (going down the rows of panels), and waiting tends to disappear in favor of pressing towards the side of the arena the subject is in, largely disregarding the stimulus in the last row.

However, it is possible that 'inertia' or a tendency to keep pressing in the direction where the subject has been pressing recently, could, in principle, completely account for this effect. In order to reject this hypothesis, one possibility would be to constrain P(L,W,R|E) conditional on a subject's history of pressing being the same. However, we do not have enough data for

such an analysis. In lieu, we look for instances where subjects have not been pressing for some time. We set this minimum refractory period preceding an action to 0.2 seconds, since this quantity is just bigger than the peak inter-press-interval across all subjects (Figure 4.13). With this constraint, it no longer is possible to compute the histogram of P(L,W,R|E) at different locations and time for each subject, but given the similarity of responses (Figure 4.25) and the histograms computed in (Figure 4.26, 4.27, 4.28, and 4.29) we collapsed the data across all subjects and compute P(L,W,R|E) as before but with the minimum 0.2 second refractory period preceding actions constraint.

In (Figure 4.30) we attempt to investigate separately the effect of time passing and of moving in the arena on the subjects' P(L,W,R|E). In each panel, the y axis represents from top to bottom: Pressing Right, Waiting, and Pressing Left. The x axis represents evidence (which goes from -1 for left to +1 for right), which we define as the mean of the stimulus for the window specified by parameters the history (N) and exclude (M) previously defined. Each column of panels represents a restricted range of positions in the arena ordered from leftmost to rightmost accordingly. Time intervals advance going down the rows of panels. First to explore the effect of time, we focus on the top row center panel, and observe what happens as we go down the rows of panels. The top row (early in the trial) has a lot of waiting across all evidence values, with maximum waiting at 0 evidence. There are a few responses to the right and left sides, which depend on the evidence, occurring mainly at 1 for right, and at -1 for left. As we move down rows of panels, waiting gets more faint, and disappears for the bottom row. In the last row, waiting is nearly completely replaced by pressing to the Right and to the left. The same observation of waiting decreasing and disappearing by the last row is seen at the off-center columns, with the notable change that the side of pressing that replaces waiting is the same as the side of the arena. We then focus on spatial signatures of commitment by focusing on one row of panels at a time and moving across the columns. For all rows, moving across the panels of columns shows us that space too has the effect of committing the subjects towards the side that they are in. The pattern is much stronger for the last row than for the first row, showing the interaction of space and time in generating the positive feedback that we call commitment.

We cannot differentiate the spatial vs the temporal dependence of commitment from this analysis, however, we in the next section we investigate

whether time and space dependant bounds are necessary to observe commitment as described in this section.

This analysis also gives us insight about the biases of the subjects. For the center column, one would expect equal probabilities of left and right responses from an unbiased subject, however our super subject has a slight left side bias. This was expected as subjects 3 and 4 have strong biases to the left (Figure 4.28) and right (Figure 4.29) respectively, which appear to cancel each other out. However, subject 1 appears to have a left bias (Figure 4.26) smaller than those of subject 3 and 4, but larger than subject 2's slight right bias (Figure 4.27). For the super subject, equal probabilities of left and right occur for the first right side positions. The psychometrics from 3 subjects show a small left stimulus response bias, and the other has a small right side bias (Figure 4.7), which result in a small left side bias for the average psychometric (Figure 4.8). The shifted psychometric suggests the left side bias observed in the action space histogram for the super subject (Figure 4.30) is a sensory bias, but there remains the possibility that this bias is a motor bias resulting from it being on average easier to press left than right for our subjects.

The commitment effect of time and space are observable in our subjects' behavior, even after considering only responses to have a minimum 200 millisecond refractory period preceding them. The fact that the precise inter-press-interval is tuned to the difficulty of the stimulus (Figure 4.17) is evidence that the actions we observe in the arena are mindful sequences of mini-decisions. These histograms are evidence that time and space affect the sequences of mini-decisions, ultimately biasing the overall categorical choice being made. In the final sections we show preliminary work investigating whether the feedback loops observed in the subjects are caused by the time and space costs in our agent's value function (3.3.3). Alternatively, the statistics of the stimulus could be the origin of the positive feedback.

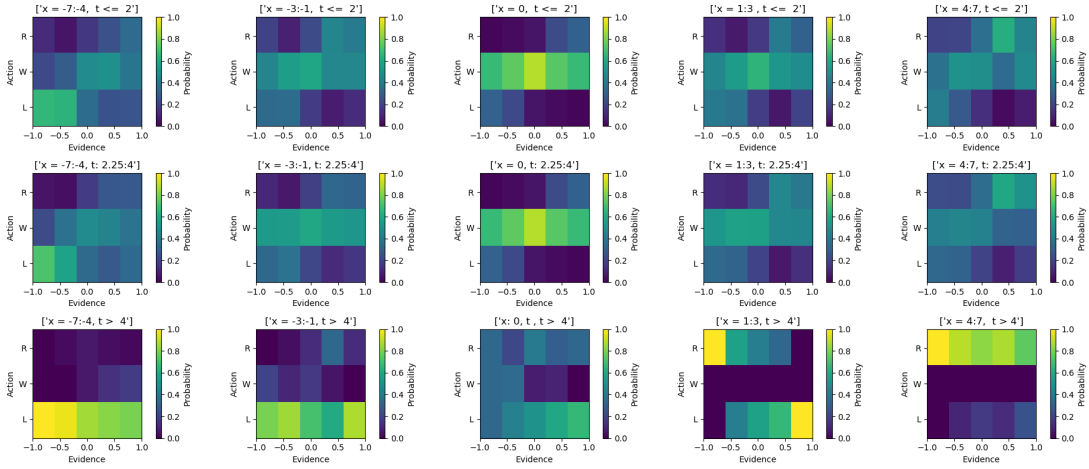


Figure 4.30: P(L,W,R|E) at different locations and time intervals in the arena for all subjects collapsed, with a minimum refractory period of 0.2 seconds between unique presses. For each row, the leftmost (rightmost) plots represent the leftmost (rightmost) spaces in the arena, and the center plot is the center of the arena. Time is split from 0 to 2 seconds (top row), 2 to 4 seconds (middle row), and 4 to 7 seconds (bottom row). At each time step, we count the number of times the subject waited, moved left, or moved right within the next 0.25 seconds, given how much evidence the subject has experienced. Evidence is defined as 3.3.4.3 over the last N samples, excluding the latest M to account for the delay between experiencing a stimulus and making a deliberate press. Results are shown for N = 7 and M = 3.

4.4.4.5 Commitment in the Optimal Agent

In this and the next section, we next conduct the same type of analysis as in (Figure 4.30) of the probabilities of pressing left, waiting, and pressing right for the optimal agent and a null model with constant decision bounds at all times and spaces in order to investigate the origin of the signature of commitment. We show that this commitment is, not surprisingly, also observed in the optimal model at the level of behavior. In chapter 3, we inferred some properties about the action space of an optimal RL agent by studying her value function. For example, we were able to abstract decision bounds for the agent at every space and time in the arena, and show that these bounds widen as the cost of time passing is relaxed (equivalent to

longer arenas), and terminate earlier the wider the arena size (fatter arenas). The optimal agent has a commitment policy by construction, since time and movement costs are imposed in the agent's value function, and these control the spatial and time properties of the decision bounds (Figure 3.18 and 3.23).

The analysis is as in the subjects' analysis (Figure 4.30). However, in the model analysis the constraint that the agent has not moved for a certain number of intervals is removed. Note that this was chosen to be 200 ms in the subject data analysis (smaller than the peak inter-press-intervals (Figure 4.13), and is also smaller than the 250 ms step size, so this constraint would bear zero effect on this analysis for the model. We set a window of 4 samples to match the window used for the subjects, however setting N to the maximum value of 28 does not qualitatively change the plots, likely because the agent has a perfect memory. The exclude parameter is set to 0 since the agent immediately perceives and integrates the stimulus. We simulated 30,000 trials for this analysis.

We see a similar pattern as in the subject behavior (Figure 4.30), for each column of panels (region of space) we can see that going down the rows (advancing in time) shifts the distribution of actions from waiting (within each panel, the middle row) to pressing left or right according to the side that the agent finds herself in (left for the leftmost panels, and right for the rightmost panels). This shows that there is a time dependency in the action space of the agent in response to the stimulus, as we expected from the decision bounds. Likewise, focusing on one row and moving down the columns we can see a shift in the probability of pressing from left to right, as the agent moves from left to right positions in the arena. Unlike the subjects, the super subject, the agent is not biased, best observed for the last row, center panel. We believe that the pattern of commitment in time observed in the subject data (Figure 4.30) and in the normative agent (Figure 4.31) is due to the costs of time (explicit in both the task and the model) and the cost of moving (implicit in the task from the geometry and the time it takes to move, explicit at each time-step in the model). In order to test this hypothesis, in the last results section (Figure 4.32 and 4.33), we begin to explore simpler null policies to generate behavior, such as flat bounds that do not change with time nor with location.

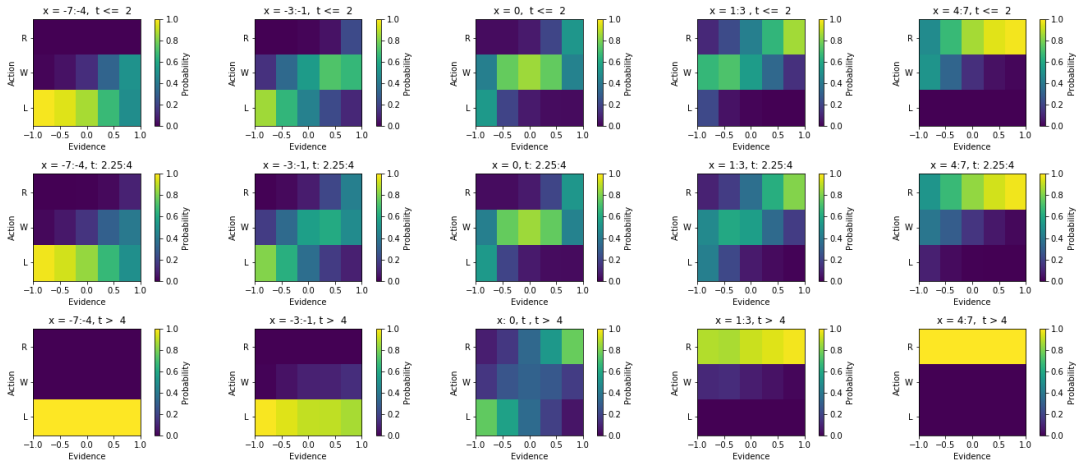


Figure 4.31: P(L,W,R|E) at different locations and time intervals in the arena for 30k simulated trials of the optimal RL agent. For each row, the leftmost (rightmost) plots represent the leftmost (rightmost) spaces in the arena, and the center plot is the center of the arena. Time is split from 0 to 2 seconds (top row), 2 to 4 seconds (middle row), and 4 to 7 seconds (bottom row). At each time step, we count the number of times the subject waited, moved left, or moved right within the next 0.25 seconds, given how much evidence the subject has experienced. Evidence is defined as 3.3.4.3 over the last N samples, excluding the latest M to account for the delay between experiencing a stimulus and making a deliberate press. Results are shown for $N = 4$ and $M = 0$, since the agent has perfect memory.

4.4.4.6 Flat decision bounds reproduce some but not all the signatures of commitment

In order to test whether the time and space dependent decision bounds of our optimal agent are necessary to produce the phenomenology of commitment towards the side the agent is already in, we simulate the exact same optimal agent as in (Figure 4.31) and use the bounds she encounters at the 3rd (18th) time step as an example of a wide (narrow) decision bound. We fix this constant bound at all times and positions to create a null model where we can observe the amount of commitment that arises naturally in our task as a consequence of guiding an agent with a stochastic stimulus, without the explicit specification of a cost on time or movement.

In order to argue that neither wide flat bounds (Figure 4.32) nor nar-

row flat bounds (Figure 4.33) produce the strong time dependent shifts in P(L,W,R|E). We find it useful to separate the discussion of "space dependence" of the probabilities from "time dependence", although these concepts may not actually have differential effects. In what concerns time, in some columns - see outermost columns of (Figure 4.32 and 4.33) for example - there in fact appears to be an increase in the probability of waiting as time goes by (moving down rows of panels), which is the exact opposite of what we observed in the subjects (Figure 4.30) and in the optimal agent (Figure 4.31). Regarding space dependence, there are some differences as a function of the position the agent is in that depend on the width of the bounds. For the wide flat bounds (Figure 4.32), the decrease in the probabilities of waiting as one moves from the center panel to the 2nd and 4th columns and as one moves from columns 2 and 4 to columns 1 and 5, respectively is much lower than for the subjects (Figure 4.30). For the narrow flat bounds (Figure 4.33), we have overall even less space-dependence of the probabilities. Since the bounds are narrow and it is therefore easy to cross bounds to either side, the probability of waiting is low everywhere. This attenuates any space dependent differences that could possibly exist. The space dependent differences that remain clear (i.e. in the top row, comparing the two extreme columns with the three middle positions) are clearly not dependant on the space and movement costs. Since the observable differences between the optimal and flat bound policies at the level of behavior may be quantitatively small, it may be fruitful to quantify the actual differences between the P(L,W,R|E) for the different policies. Visual inspection suggests removing the time and space dependence of the decision bounds leads to a decrease in commitment behavior.

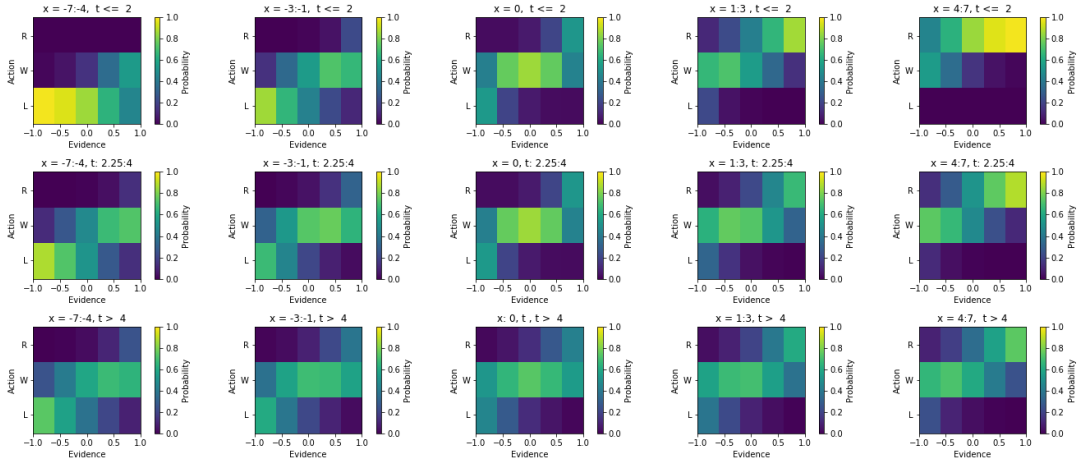


Figure 4.32: P(L,W,R|E) at different locations and time intervals in the arena for 30k simulated trials of an agent with wide flat action bounds that do not change in space nor time. For each row, the leftmost (rightmost) plots represent the leftmost (rightmost) spaces in the arena, and the center plot is the center of the arena. Time is split from 0 to 2 seconds (top row), 2 to 4 seconds (middle row), and 4 to 7 seconds (bottom row). At each time step, we count the number of times the subject waited, moved left, or moved right within the next 0.25 seconds, given how much evidence the subject has experienced. Evidence is defined as 3.3.4.3 over the last N samples, excluding the latest M to account for the delay between experiencing a stimulus and making a deliberate press. Results are shown for N = 4 and M = 0, since the agent has perfect memory. We set a window of 4 samples to match the window used for the subjects, however setting N to the maximum value of 28 does not qualitatively change the plots. We simulated 30,000 trials with bound height comparable to the bounds observed by the optimal agent of (Figure 4.31) at the 3rd time-step in the center of the arena.

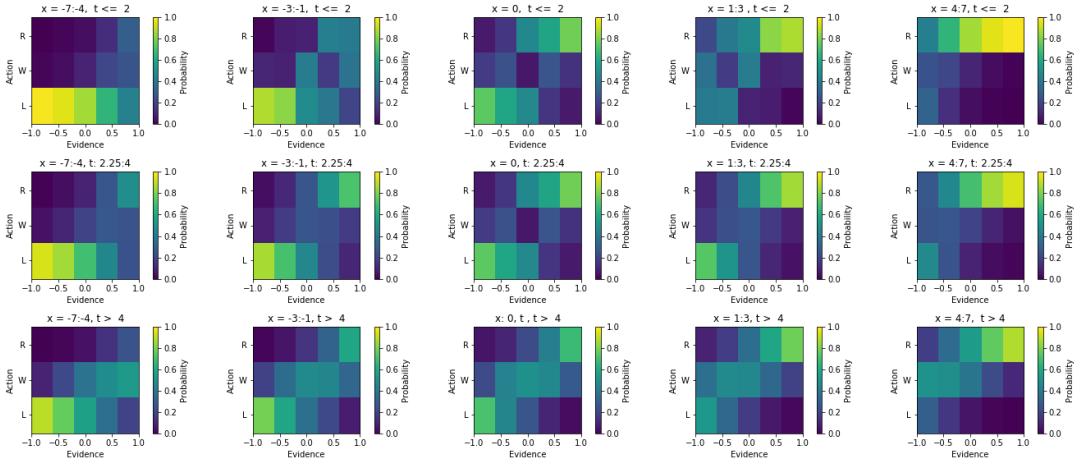


Figure 4.33: P(L,W,R|E) at different locations and time intervals in the arena for 30k simulated trials of an agent with narrow flat action bounds that do not change in space nor time. For each row, the leftmost (rightmost) plots represent the leftmost (rightmost) spaces in the arena, and the center plot is the center of the arena. Time is split from 0 to 2 seconds (top row), 2 to 4 seconds (middle row), and 4 to 7 seconds (bottom row). At each time step, we count the number of times the subject waited, moved left, or moved right within the next 0.25 seconds, given how much evidence the subject has experienced. Evidence is defined as 3.3.4.3 over the last N samples, excluding the latest M to account for the delay between experiencing a stimulus and making a deliberate press. Results are shown for N = 4 and M = 0, since the agent has perfect memory. We set a window of 4 samples to match the window used for the subjects, however setting N to the maximum value of 28 does not qualitatively change the plots. We simulated 30,000 trials with bound height comparable to the bounds observed by the optimal agent of (Figure 4.31) at the 18th time-step in the center of the arena.

Regardless of whether the bound applied is wide (Figure 4.32) or narrow (Figure 4.33), even late into the trial the agent continues to wait and move in the opposite direction than the target it is closest to. Critically, both of these behaviors are reduced as time advances in the subject data and in the optimal RL agent. The future steps for this work will involve quantifying the degree to which these probabilities differ.

4.5 Discussion

In this chapter we managed to capture hesitant choice behavior by extending a standard 2AFC task to a 2D arena that must be traversed in order to report a binary choice. Recent studies are similarly beginning to analyze behavioral data during the execution of a task, intuiting that it is a meaningful part of the cognitive cost function [Spivey et al., 2005, Resulaj et al., 2009, Haith et al., 2015, Selen et al., 2012, Wong and Haith, 2017]. The nature of the questions posed by [Haith et al., 2015] and their approach is strikingly close to our study. They investigate experimentally the range of distances between targets for which in-between trajectories occur in a 'go-before-you-know' task, and then model a normative agent that captures the dependence of in-between trajectories on the geometric cost imposed. The authors explicitly say that their objective was not to fit the data to a model, but to show that a normative model with explicit goals that depend not only on stimulus probabilities, but also on motor constraints from the environment could produce human-like behavior, suggesting such trajectories may not be errors due to competing motor plans, but rather deliberate actions for dealing with an uncertain world [Haith et al., 2015, Wong and Haith, 2017]. We go further in suggesting that in-between movements are deliberate, and provide several pieces of evidence that intermediate actions are deliberate, and dependant on the temporal characteristics of the stimulus nearly on a moment-to-moment basis (Figure 4.3, 4.22, and 4.25). The overall results suggest that the immediate next action of subjects is significantly linearly correlated with the count of the stimulus at the previous time-step, and that this correlation is strongest at the beginning of the trial. We found it important to use a time-varying stimulus, as it allowed others to make inferences about the time dependence of choice behavior on the stimulus [Nienborg and Cumming, 2007, Brunton et al., 2013]. We were inspired early on by a study that used time-varying visual stimuli to assess via psychophysical reverse-correlation which periods were most informative to choices for which the mean stimulus signal was weak [Nienborg and Cumming, 2007]. Drawing from all of our experience across 3 experiments and the normative model we form some conclusions. In both cases, the not fully committed state of the agent or subject must be in a sweet spot of time urgency and movement compatibility. Hesitation behavior defined as the Actual Number of Presses divided by the Minimum Required Number of Presses is seen to be maximum in the model

for arenas that are smaller than 20 grids, and conversely the cost of time cannot be too small (Figure 3.24). In experiment 1, we didn't have any time limits beyond the maximum time allowed of 8 seconds. This put subjects in the regime where waiting until one is fully certain to act is the most advantageous thing to do, and so only a few hesitant decision trajectories were captured. While we managed to put subjects in Experiment 2 in a regime of higher cost of time, similar as [Haith et al., 2015] we didn't have a precise grip on this tie urgency, and our space state representation is a gross simplification of the motor component in the human task. The grid world allowed us to explore regimes of geometric cost for the human task in comparable terms to the model. Pilot experiments were used to optimise for the arena size to 8 steps to each side (17x) and 7 seconds time limit, with points being lost linearly from 100 to 0 in that time. This resulted in behavior qualitatively similar to what we saw in the model, where we understand the reason why similar stimuli result in similar trajectories. However, this similarity is difficult to show in the human data. We analyzed the behavior of the subjects depending on their position in the arena, and found signatures of commitment suggested by the model. It is also possible to make such plots for the model, which not surprisingly show the same commitment patterns. In order to test if the pattern of commitment in time observed in the subject data (Figure 4.30) and in the normative agent (Figure 4.31) is caused by the costs of time and the cost of moving, we began to explore simpler null policies to generate behavior, such as flat decision bounds that are the same at all times and positions. Preliminary results suggest that commitment, as described in the last sections, is not fully dependant on the time and movement costs (Figure 4.31), since some of the signatures of commitment we observed in the optimal agent are also present for both the wide bounds (Figure 4.32) as well as for narrow bounds (Figure 4.33), however not as pronounced. To observe spatial effects of commitment, we focus on one row at a time and can see that P(L,W,R|E) changes for the flat bound models as we move across columns (positions in the arena), even if these changes are not as strong as in the optimal model or the subject data. In contrast, the behavior of the flat models with respect to time (moving down the rows of panels) clearly is different to the optimal agent (Figure 4.31) and the subjects (Figure 4.30). Unlike the subjects or the optimal model, the flat bound model continues to wait or even waits more late into the trials for both the wide bounds (Figure 4.32) as well as for narrow bounds (Figure

4.33). A quantification of the degree to which P(L,W,R|E) changes in the optimal policy relative to the flat bounds policy might reveal more about the dependence of this change on the time and movement costs, or whether this decrease is due to the behavior of the belief function on the stimulus itself. One could explore the dependence of commitment on the stimulus by constraining P(L,W,R) on the value of belief (which is what directly makes the flat bounds agent move) instead of the evidence. In the top row of all the models (flat and optimal), we see a bias for continuing to go in the same direction when the agent is located in the more extreme positions (columns 1 vs 5). This can be seen as a form of confirmation bias, but unlike the confirmation bias that arises in active sensing, where the observations that follow are affected by the organismâĂŹs actions, actions do not alter the time-course of our stimulus. However, part of the commitment is still expected, because it reflects that stimuli have some signal which is maintained through the stimulus. However, this doesn't explain the commitment we observe fully, since we are constraining these histograms on a range of evidence. One might be able to factor the expected bias of the stimulus by constraining on the exact belief path. This analysis is likely only testable in our simulated agents due to the large amount of data such a test would require.

4.6 Acknowledgments

Chapter 4 was analyzed by me with the guidance of Alfonso Renart. We would like to thank Miguel Bengala for adapting the commitment analysis for the RL agent and flat action bounds agent. We would like to thank Gautam Agarwal, Juan Castiñeiras, and Zach Mainen for discussions.

Bibliography

[Brunton et al., 2013] Brunton, B. W., Botvinick, M. M., and Brody, C. D. (2013). Rats and humans can optimally accumulate evidence for decision-making. *Science (New York, N.Y.)*, 340(6128):95–98. (Cited on pages 54 and 145.)

[Drugowitsch et al., 2012] Drugowitsch, J., Moreno-Bote, R. N., Churchland, A. K., Shadlen, M. N., and Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. *Journal of Neuroscience*. (Cited on pages 6, 7, 9, 53, 58, 59, 60, 62, 65, 66, 74, 76, 78, 97, 98, 106, 153, and 154.)

[Haith et al., 2015] Haith, A. M., Huberdeau, D. M., and Krakauer, J. W. (2015). Hedging Your Bets: Intermediate Movements as Optimal Behavior in the Context of an Incomplete Decision. *PLoS Computational Biology*. (Cited on pages 97, 98, 145, 146, and 152.)

[Nienborg and Cumming, 2007] Nienborg, H. and Cumming, B. G. (2007). Psychophysically measured task strategy for disparity discrimination is reflected in V2 neurons. *Nature Neuroscience*. (Cited on pages 34, 52, and 145.)

[Resulaj et al., 2009] Resulaj, A., Kiani, R., Wolpert, D. M., and Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature*. (Cited on pages 8, 9, 51, 52, 145, 151, and 152.)

[Selen et al., 2012] Selen, L. P., Shadlen, M. N., and Wolpert, D. M. (2012). Deliberation in the motor system: Reflex gains track evolving evidence leading to a decision. *Journal of Neuroscience*. (Cited on pages 9, 20, 27, 34, 52, 145, and 151.)

[Spivey et al., 2005] Spivey, M. J., Grosjean, M., and Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences of the United States of America*. (Cited on pages 8, 20, 27, 34, 51, 52, 145, 151, and 152.)

[Wong and Haith, 2017] Wong, A. L. and Haith, A. M. (2017). Motor planning flexibly optimizes performance under uncertainty about task goals. *Nature Communications*. (Cited on pages 98 and 145.)

Chapter 5

Final Discussion

HIGHLIGHTS

- In this chapter we summarize all of our results and discuss them in the context of the decision-making literature.

Contents

5.1 Final Discussion 151

5.1 Final Discussion

We introduced a novel psychophysics task in order to measure behavior during the deliberation period of a binary choice task. Similar methods had been used previously to investigate decision making processes from motor output [Spivey et al., 2005, Resulaj et al., 2009, Selen et al., 2012], however these investigations used either very simple stimuli [Spivey et al., 2005], or extinguished the stimulus when movements began [Resulaj et al., 2009, Selen et al., 2012]. Our investigation was unique in that we captured motor output while subjects experienced a stochastic, time-varying stimulus.

Our novel experiments allowed us to capture in-between trajectories in a fraction of the hard trials of Experiment 1 (Figure 2.7), however, this phenomenology happened too few times to statistically assess its connection to the stimulus. In order to increase the number of trajectories per experiment,

we imposed a strict time limit of half a second to initiate movements. Similar constraints were utilized in previous investigations of the in-between movements [Resulaj et al., 2009, Hanks et al., 2015]. This modification resulted in a lot more complex trajectories per experimental session (Figure 2.15 and 2.17). This phenomenology can be abstracted to a simple summary statistic, the trajectory length, which increases for all subjects as the difficulty of the stimulus increases (Figure 2.9 and 2.18). That the in-between trajectories occur for more difficult stimuli had been previously described by [Spivey et al., 2005], also using auditory stimuli (words). We characterize this phenomenon more thoroughly by systematically testing different stimulus difficulties dynamically varying the mean inter-aural level difference of white noise stimuli delivered via headphones. The key point that experiment 2 data allowed us to make is that for difficult trials, it is possible to train a ridge regression model to reconstruct the noisy stimulus sequences from linear combinations of eye and hand data (Figure 2.21 and 2.22). Above chance correlation between the reconstructions and the true trajectories is observed on hold-out cross-validation data. This is empirical evidence that the online trajectories hold information about temporal features of the stochastic stimulus (Figure 2.22 and 2.23), however it remained difficult to describe the exact role of the stimulus in guiding subject movements.

One hypothesis is that our hesitant trajectories were motor errors resulting from competing motor plans [Haith et al., 2015], but this theory would explain only erroneous movement initiation, whereas we sometimes observe several corrections within a single trial. Subjects appear to be handling the sequential sampling problem, partly, by making sequences of mini-decisions along the way. The bounded accumulation of evidence had been expanded to explain changes of mind, action corrections following an initial movement towards the wrong target [Resulaj et al., 2009]. It would take 7 more years for that framework to be generalized to incorporate 2 sequential decisions [Van Den Berg et al., 2016]. In the interim, the dependence of changes of mind on motor effort, quantified with force fields as well as varying distances between targets was demonstrated and modeled as changes to the change of mind bound (a second bound following the initial decision bound) [Burk et al., 2014]. This is qualitatively the same result we would obtain around 2016 by varying arena width in our normative agent, as well as [Haith et al., 2015] a year earlier.

Prior to our own modeling exercise, the reason why we became inter-

ested in normative models and reinforcement learning is that it provides both an output behavior and a theory of why that behavior is generated. We realized we didn't have a means of modeling optimal or even reasonable behavior in our task because we lacked a self consistent task description. Our subjects were reaching for targets merely because we instructed them to. In order to know what behavior could be expected in our embodied decision making task, we would have to describe the precise goals of subjects and agents in the task in terms of rewards, and any relevant environmental constraints in terms of costs [Sutton and Barto, 2018]. Normative solutions to the cognitive 2AFC task had been previously studied [Moreno-Bote, 2010, Drugowitsch et al., 2012]. The framework of [Drugowitsch et al., 2012] used dynamic programming to estimate the value of each state in a trial by back-propagating it from known future values of reward. The agent evaluates the probability of obtaining the rewards using Bayesian statistics, and can weight the size of that reward, R, by the probability of obtaining it, $p(R)R$, against the cost of moving and the cost for waiting to experience more stimuli and become more certain. The resulting policy imposes optimal collapsing bounds on the accumulated evidence. We generalized this previously described POMDP framework ([Drugowitsch et al., 2012]) to also include a geometric cost for approaching distally separated targets in an arena. Our agent produces hesitant decision trajectories for difficult stimuli (Figure 3.10) for a specific range of time and movement costs.

Our optimal agent produces similar behavioral summary statistics as we saw in our first two experiments, namely that the reaction times and the trajectory lengths increase significantly with increasing stimulus difficulty (decreasing mean ILD) (Figure 3.10). We characterize the number of presses and other signatures of behavior (i.e. reaction time, start time, and post-start waiting) for an agent that must traverse an arena of width 17 (unit grids) under a range of different time limits (Figure 3.17) and find that as the cost of time increases, reaction times and start times systematically decrease and approach their theoretical minimum values. The opposite effect is observed on hesitation, defined as Number of Presses/Required and post-start waits, which increase as the cost of time is increased, but also plateau and appear to decrease from some maximum value around 0.04 cost per unit time (equivalent to 7 seconds time limit for a unit reward in the disembodied task). We were able to abstract this behavior to action bounds on the beliefs of the agent, and observed that the width separating the bounds at each time

step decreases as the cost of time is increased (Figure 3.18).

Stimulus dependent reaction times due to time collapsing decision bounds was described for the optimal disembodied agent [Drugowitsch et al., 2012], and is consistent with our disembodied agent's behavior as well (Figure 3.5). However, critically for us these bounds are generally not terminal bounds until the agent finds herself with just enough time to reach a target. We call non-terminal action bounds soft decision bounds, and the cost of time controls the width between them. In general, higher time costs (equivalent to shorter time limits) create conditions in which less extreme beliefs are required for the agent to move rather than wait. For very high costs of time, the action bounds on belief approach chance level (0.5). Fixing the time limit and instead varying the arena width allowed us to investigate the effect of arena width on the behavior of the agent (Figure 3.22), and on the action bounds (Figure 3.23). At the level of behavior, increasing arena width initially leads to little to no change in the percentage of correct responses, however, if performance were determined according to the mean of the experienced stimulus, an initial improvement for small arena widths (5-11 grids) can be observed (Figure 3.6), meaning the agent is over confident. Reaction times systematically increase with increasing arena width, start times initially decrease for arenas smaller than 20 grids in width, but then plateau before decreasing again for the largest arena sizes. Average number of presses normalized by arena width initially increases, reaching a maximum value around width 11, and decreasing slowly with width 21 still close to the peak. A similar pattern but with lower decay rate can be seen in post-start waits. The non-monoticity of the signatures of behavior as a function of time and space highlight the fact that the uncommitted regime of behavior we have been interested in studying in this thesis requires very specific environmental settings of time and movement costs (arena sizes).

In order to observe hesitation in an embodied categorical choice, we have learned that experimental conditions must be precisely fine-tuned. This can explain why our first experiment did not result in many hesitant decision trajectories. We were sampling in the asymptotic behavior of the arena (long), where the agents have a lot of time to become certain before acting. Experiment 2 had more hesitant decision trajectories, but the drastic reduction in reaction times and the idiosyncrasy of the decision trajectories made us suspect that the maximum start time constraint we imposed put subjects in excess of urgency. This regime might be similar to setting the arena

size too wide. These models' signatures of behavior (Figure 3.24) gave us a range of arena sizes to explore in our final psychophysics experiments. Pilot experiments (not shown) were used to test a range of arena sizes. We found that already from the first session subjects showed a lot of hesitant behavior for arena sizes 15, 17, and 19, in line with the modeling results showing maximum hesitation for these arena sizes (Figure 3.24). We fixed the arena size at 17 for our final experiment. Very small arena widths (e.g. 5, 7) did not result in much hesitation in the subjects, but the model still produces some hesitation. This might be indicative of a motor cost for starting to move being present in the subjects.

In the optimal agent, at the level of the bounds we observed that increasing the width of the arena shifts the agent's decision bound collapse to earlier times. This is a hard bound that determines the final time for which it is still possible to reach a target before the time limit. Shifting the hard decision bounds to earlier times by definition commits the agent to a choice earlier (Figure 3.23). The effect of varying arena size on the in-between decision trajectories of human subjects had previously been described by [Hanks et al., 2015] and a similar effect on the changes of mind phenomenology had been described by [Burk et al., 2014]. The authors of [Hanks et al., 2015] in fact modeled the behavior with a reinforcement learning agent that penalizes movement in a common currency as task rewards, similar to ours. However, to our knowledge, the sequential nature of the decision guided actions that affect one another is not explicitly investigated by any decision making study prior to ours, until the publication in 2016 of the 2-stage decision study of [Van Den Berg et al., 2016]. Even in consideration of that work [Van Den Berg et al., 2016], our study better captures the sequential and continuous nature of decision making in both experimental design and theoretical approach.

Our final human experiment implements a time dependent point system in order to mimic the time costs in our model, and replaces the joystick control with button presses, negating the need for modeling high dimensional aspects of moving an arm. As in our previous tasks and our model, average trajectory length increases significantly with increasing stimulus difficulty (decreasing mean ILD) (Figure 4.11 and 4.12). This dependence on the stimulus difficulty leaks onto finer motor correlates of displacement, such as the inter-press-interval which decreases as the stimulus becomes easier (Figure 4.15). That the precise timing of button presses depends on the stimulus

can be seen in the reliability of the trajectories to repeated presentations of the same stimulus. Responses to repeated presentations of the same stimulus result in similar responses on different trials, which is true within subject (Figure 4.23) as well as across subjects (Figure 4.25) and also depends on the difficulty of the stimulus (Figure 4.22 and 4.24).

While our model explicitly makes a choice at every time step, we do not think that this is what our human subjects are doing, at least not at every time-step. We do however believe that the hesitant decisions reflect sequences of decisions preceding, at least, at every change of direction. Sequential decision making typically describes 'a situation where the decision maker makes successive observations of a process before a final decision is made' [Diederich, 2015], which is the situation that standard drift diffusion provides a solution to. In contrast, our investigation has more in common with the field of dynamic decision making, which describes a 'situation where a series of actions must be taken over time to achieve some overall goal; the actions are interdependent so that later decisions depend on earlier actions; and the environment changes both spontaneously and as a consequence of earlier actions, realistic examples include fighting fires, navigational control, battlefield decisions, medical emergencies, and so on' [Busemeyer, 2001]. In these examples, the dependence of future states on past states is typically by construction (e.g the state of a chess board after a few moves, or the state of a fire after fighting it from a few angles). The dependence on each action on the previous is also by construction, but not explicitly. In our study, the available actions at each time step are identical, as is the way the stimulus is processed (always Bayesian statistics). The dependence in space is rather a result of the iterative derivation the past values from expected future rewards and action costs.

Whereas complex behavior typically arises in dynamic decision making tasks out of complex interactions with the environment, in our study hesitant stimulus correlated choice trajectories naturally arise during the deliberation of simple perceptual stimuli, seemingly out of simple temporal and spatial constraints. We also highlight that although we are concerned with motor aspects of a decision, our investigation is not concerned with the actual control of the arm. We purposely do not attempt to model the cost of operating biomechanically realistic limbs, which is a difficult problem tackled by optimal feedback control [McNamee and Wolpert, 2019, Todorov and Jordan, 2002]. Instead, we focus on understanding a more cognitive aspect of control related

to taking into consideration the geometry of the environment that must be navigated, and the time that it takes to carry out actions. We reduce the action space to simple binary actions in order to mitigate the importance of real bio-mechanical costs in our study. Our simulations show that, for a range of time and movement cost regimes, hesitant, in-between decision trajectories are optimal. The model makes a prediction that not only time, but also space can commit an optimal agent to decide.

We looked for signatures of commitment in subject behavior and find that P(L,W,R|E), the probability of pressing left, right, or waiting given the evidence is dependant on the position of the subjects in the arena and time in the task (Figure 4.30), as well as in the optimal model (Figure 4.31).

In both the subjects and the optimal agent, focusing on the top row of panels and looking across the columns, we can see that the position of the subjects (the column) affects P(L,W,R|E), with the outermost columns having maximum probabilities of any action for presses towards the side that the subject is on. This pattern is more easily observed as one looks down the rows of panels. This shows a time dependence on the commitment phenomenology we describe, suggesting that the effect of time and space on this form of positive feedback towards the side the agent is interdependent on time and space. We have looked for these signatures of commitment in simpler null models of behavior, such as flat decision bounds at all space and times to investigate if the origin of commitment is due to the time and space collapsing bounds (Figure 4.32 and 4.33). Our preliminary results suggest that time related signatures of commitment are not present, or even reverse in the flat bounds model, however space related signatures are still observable in the flat bound model.

We introduced a new task and a theoretical framework that extends the 2AFC task by incorporating a spatial navigation component. Our task allows us to measure a more continuous output of decision behavior in the 2AFC task, during the deliberation period, revealing an uncommitted behavior that we call "hesitation". The hesitant decision trajectories occur in our theoretical framework only for very specific environments (e.g. hesitation is not seen for very long or very wide arenas). Our first two experiments gave us a (painful) preview of the finding that hesitation only occurs for finely tuned environmental parameters.

We believe that our consideration of extended actions as opposed to a single, binary event is critical to understand the nature of commitment. We

found evidence that commitment is not actually all or nothing, but that an uncommitted state can be observed in situations requiring agents to navigate an environment to report a binary choice. However, the hesitant trajectories we observed are rare. In fact we learned that hesitant behavior only occurs for specific parameters of arena width and arena height (time limit). We developed a task and a theoretical framework where we can control the factors affecting extended decisions (time and space). This control over the urgency to act was critical to find the hesitant regime. Although computing the optimal policies is a difficult endeavour, we find that the behavior generated by our simulated agents allows us to further study the origins of hesitation. For example, we can compare the hesitation in the optimal agent (which is under our control), to that in the subjects (which is to some degree under our control), to, for example, the behavior of agents lacking any change in their decision bounds (flat bounds). Some limitations of our optimal agents, aside from the effort to model them, are that tens of thousands of trials are needed to produce informative plots, and validating such work in human subjects may thus be impossible. Furthermore, there are some discrepancies between the mode of pressing in the subjects compared to the agents. For example, for very small arenas subjects do not engage in hesitant behavior (pilot experiments not shown), whereas the agent hesitates even for small arenas. Additionally, the subjects have a tendency to continue pressing to the same side they are already pressing, which can be so fast (<200 ms inter-press-interval) that it is more likely the result of some motor bias to continue pressing than due to the type of cognitive commitment we are studying in the optimal agent. It is very likely that the human subjects employ completely different, simpler heuristics for solving the task, than perfect integration of evidence and optimal trade-offs. Furthermore, strategies might differ across subjects. Our final results testing simpler action policies suggest that to some degree, flat bounded accumulation of evidence can account for the commitment effects we observed, more so in space than time. This will be further investigated in the lab, first by quantifying the amount to which the spatial and temporal signatures of commitment change by the removal of the spatial and temporal bound collapse.

Taken together, the experimental and theoretical work in this thesis is a conceptual bridge between the fields of sequential decision making and dynamic decision making, and is therefore a meaningful contribution to the broader field of perceptual decision making and to neuroscience.

Bibliography

[Burk et al., 2014] Burk, D., Ingram, J., Franklin, D., Shadlen, M., and Wolpert, D. (2014). Motor Effort Alters Changes of Mind in Sensorimotor Decision Making. *PloS one*, 9:e92681. (Cited on pages 98, 152, and 155.)

[Busemeyer, 2001] Busemeyer, J. R. (2001). Dynamic Decision Making. In Smelser, N. J. and Baltes, P. B., editors, *International Encyclopedia of the Social & Behavioral Sciences*, pages 3903–3908. Pergamon, Oxford. (Cited on page 156.)

[Diederich, 2015] Diederich, A. (2015). Decision and Choice: Sequential Decision Making. In *International Encyclopedia of the Social & Behavioral Sciences: Second Edition*. (Cited on page 156.)

[Drugowitsch et al., 2012] Drugowitsch, J., Moreno-Bote, R. N., Churchland, A. K., Shadlen, M. N., and Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. *Journal of Neuroscience*. (Cited on pages 6, 7, 9, 53, 58, 59, 60, 62, 65, 66, 74, 76, 78, 97, 98, 106, 153, and 154.)

[Haith et al., 2015] Haith, A. M., Huberdeau, D. M., and Krakauer, J. W. (2015). Hedging Your Bets: Intermediate Movements as Optimal Behavior in the Context of an Incomplete Decision. *PLoS Computational Biology*. (Cited on pages 97, 98, 145, 146, and 152.)

[Hanks et al., 2015] Hanks, T. D., Kopec, C. D., Brunton, B. W., Duan, C. A., Erlich, J. C., and Brody, C. D. (2015). Distinct relationships of parietal and prefrontal cortices to evidence accumulation. *Nature*. (Cited on pages 7, 152, and 155.)

[McNamee and Wolpert, 2019] McNamee, D. and Wolpert, D. M. (2019). Internal Models in Biological Control. *Annual Review of Control, Robotics, and Autonomous Systems*. (Cited on page 156.)

[Moreno-Bote, 2010] Moreno-Bote, R. (2010). Decision confidence and uncertainty in diffusion models with partially correlated neuronal integrators. (Cited on pages 9, 59, 60, and 153.)

[Resulaj et al., 2009] Resulaj, A., Kiani, R., Wolpert, D. M., and Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature*. (Cited on pages 8, 9, 51, 52, 145, 151, and 152.)

[Selen et al., 2012] Selen, L. P., Shadlen, M. N., and Wolpert, D. M. (2012). Deliberation in the motor system: Reflex gains track evolving evidence leading to a decision. *Journal of Neuroscience*. (Cited on pages 9, 20, 27, 34, 52, 145, and 151.)

[Spivey et al., 2005] Spivey, M. J., Grosjean, M., and Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences of the United States of America*. (Cited on pages 8, 20, 27, 34, 51, 52, 145, 151, and 152.)

[Sutton and Barto, 2018] Sutton, R. and Barto, A. (2018). Reinforcment Learning. *MIT Press*. (Cited on pages 58 and 153.)

[Todorov and Jordan, 2002] Todorov, E. and Jordan, M. I. (2002). Optimal feedback control as a theory of motor coordination. *Nature Neuroscience*. (Cited on page 156.)

[Van Den Berg et al., 2016] Van Den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N., and Wolpert, D. M. (2016). A common mechanism underlies changes of mind about decisions and confidence. *eLife*. (Cited on pages 98, 152, and 155.)

Apoio financeiro da FCT e do FSE no âmbito do Quadro Comunitário de Apoio, Bolsa nº [SFRH / BD / 51896 / 2012].

ITQB-UNL | Av. da República, 2780-157 Oeiras, Portugal
Tel (+351) 214 469 100
Fax (+351) 214 411 277

www.itqb.unl.pt