Investigating the complexity of interactions between attributions and beliefs: evidence from a novel task

Elena Zamfir

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Gatsby Computational Neuroscience Unit

University College London

May 8, 2021

I, Elena Zamfir, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

This thesis presents our work investigating the nature of and interactions between processes underlying causal attributions and the formation and updating of beliefs. The two research directions that constitute the main inspiration for our work - research in psychiatry and computational accounts of decision-making - have traditionally been separate; however there has recently been a growing effort to bridge the gap by bringing computational tools to bear on fundamental research questions in psychiatry. We contribute to this effort by designing a quantitative framework for phrasing, exploring and testing hypotheses associated with the attribution-self-representation cycle theory.

We developed a novel task, in which attributions and beliefs about the self are measured repeatedly, producing the time series data necessary to investigate interactions between these two variables on a trial-by-trial basis. Importantly, subjects' beliefs and causal attributions are probed with regard to real outcomes, experienced in the context of learning a skill task, and in the absence of any manipulation targeting their content.

We present evidence of effects consistent with the cycle postulated by the theory, namely trial-level effects of attributions on beliefs about skill and effects of beliefs about skill on attributions, neither of which can be reduced to the effect of objective performance.

The richness of the task enabled the revelation of substantial behavioural complexity, suggesting testable hypotheses for future work. Of note among these are questions about the modulation of attribution-belief interactions by outcome valence, and the factors governing differential processing of various

task features.

In sum, our work proposed and implemented a novel framework for investigations into the dynamics of beliefs and causal attributions, and completed the first steps towards a precise formalisation and testing of the theoretical proposal in this framework, while providing novel evidence in support of the theory.

Impact statement

This thesis presents our work investigating the interactions between causal attributions for outcomes and the updating of beliefs about oneself. Affectively catastrophic positive feedback processes by which otherwise surmountable negative life events turn into major psychological hurdles constitute an important aspect of psychiatric dysfunctions, which are associated with huge psychological, health and economic costs (WHO, 2008). We aimed to contribute to a better understanding of these processes by designing and using a quantitative framework for phrasing, exploring and testing hypotheses associated with the attribution-self-representation cycle theory.

This thesis presents a review of ample research literature on attribution, and provides a unifying perspective on research from diverse fields, along with illustrating the use of computational approaches in data analysis and in the framing of new hypotheses. As such, it can have a direct and immediate impact by serving as a valuable resource for students and early career researchers in the field, and by encouraging similar integrative approaches.

We provide a novel task that can be used to probe attribution making and its relationship with beliefs, present new evidence in support of the theoretical proposal, and suggest testable hypotheses for future work. We therefore hope our work has a broader impact in the field, by encouraging further research on testing and developing the theory, and more widespread development and deployment of complex tasks to investigate these phenomena.

Finally and most importantly, we hope this work contributes to the pursuit of a better understanding of the dynamics of causal attribution, beliefs

about the self (and others), and their interactions. By uncovering mechanisms associated with psychiatric disorders onset and maintenance, as well as protective factors that could be used to design targeted interventions, such understanding would undoubtedly have a real and important impact in society at large.

Acknowledgements

I am very grateful for the things I learned, the people I met and the experiences I had during my PhD. It has been a very intense and enriching period, in which I feel I've grown both as a researcher and as a person.

I want to thank Peter Dayan for being an inspiring scientist and advisor, for being so generous with his time, advice and comments, and for his patience and sense of humour. I am grateful and honoured for having had the opportunity to work with him, and to get to know him as a scientist and as a person.

I also want to thank Jonathan Roiser for welcoming me so warmly to his lab, and for giving me the opportunity to experience a different atmosphere and lab culture. I learned a lot about the perspective, questions and critical approach of the experimenter from these lab meetings, and I am very grateful for this.

I want to thank everyone at Gatsby who made my time there so enjoyable and enriching. I am grateful to Peter Latham for the very first thing he said during the very first Theoretical Neuroscience course - I have, since then, had ample opportunity to appreciate even more how difficult and essential it is to find questions that are both interesting and answerable, for the TN assignments, which I've enjoyed so much and which were such a good learning experience, and for his trademark question during talks. I am grateful to Maneesh Sahani for his guidance during my minor, and for being the one who first unveiled the magic of Gaussian Processes to me, and to Arthur Gretton for his energy, enthusiasm and unusual sense of humour.

I also want to thank Barry Fong, Reign Mcmillan, Mike Sainsbury and Ana Saraiva for their permanent support and patience and for making everything run so smoothly.

And of course, I am grateful for all the interesting, smart, funny, slightly crazy, passionate, knowledgeable, generous people that made Gatsby what it meant to me: a place that became my first London home, intellectually and (almost) physically. I've learned so much from our conversations, about neuroscience and machine learning, about myself, and about life, the universe and everything. So thank you all : Ritwik Niyogi, Balaji Lakshminarayanan, Gergo Bohner, Federico Mancinelli, Sofy Jativa, Joana Soldado Magraner, Heiko Strathmann, Laurence Aitchison, Pedro Goncalves, Ben Huh, Carsen Stringer, Vincent Adam, Witttawat Jitkrittum, Zoltan Szabo, Mijung Park, Kacper Chwialkowski, Sanjeevan Ahilan, Franziska Broeker, Ricardo Monti, Kirsty McNaught, Roman Pogodin, Ilyes Khemakhem, Michel Arbel, Kevin Li.

I am also grateful for the opportunity to know and learn from people at the Sainsbury-Wellcome Center, and I want especially to thank Goncalo Lopez, Joana Nogueira, Joana Neto and Lorenza Calcaterra for welcoming me into their lab-family, for being such good friends, and for all the wonderful moments we spent together.

I want to thank everyone at reception at both ICN and SWC for welcoming me with a smile every day, and for their kindness and well-meaning help with various issues.

I've had the opportunity to meet and learn from many other people, at summer schools and conferences that I attended. I am also grateful for these experiences, which have been exciting and enlightening, and so enjoyable. I am particularly grateful for the intellectual effervescence of the FENS Chemical Neuromodulation summer school in Bertinoro in 2017 and of the Transylvanian Experimental Neuroscience Summer School in 2016.

I am also grateful to the friends not yet mentioned, who supported and

encouraged me during this period, with whom I could share both the ups and the downs of the process: Andra Juganaru, Ruxandra Cojocaru, Alina Cristina Marin, Oana Lang, Ozana Nitulescu, and to Georgiana Petria, who hosted me at the start of my London experience.

Above all, I am grateful to my family for being constantly and unconditionally supportive and loving, and to whom I owe everything.

Contents

List of Figures

List of Tables

Glossary

- **skill estimates**: subjects' estimates of their own (or the "other"'s) skill; raw values coded on a continuous scale in $[0, 1]$;
- **skill updates**: differences between two successive skill estimates;
- **session break effect**: large difference between the last skill estimate from the first session and the first one from the second session;
- **attributions**: subjects' causal attributions for the outcomes they (or the "other") just experienced; provided as a choice among the following options: $\{\text{internal}(I), \text{maze}(M), \text{rotations}(R), \text{luck}(L)\}$;
- **external attributions**: attributions to maze, rotations and luck;
- **pl**: path length; length of correct path through a maze;
- **pnu**: proportion of non up orientations; proportion of time within a trial during which the maze orientation is not the normal UP orientation;
- **pc**: proportion correct; proportion of correct key presses out of all key pressed by the subject during a trial;
- **pp**: proportion pauses; proportion of time spent not pressing any key out of all time available for a trial;
- **pwcu**: proportion wrong, correct for up; proportion of wrong key presses that would have been correct in the normal UP orientation;
- mathematical notation will be defined as it is introduced in the text.

Introduction

This thesis presents our work investigating the interactions between causal attributions for outcomes and the updating of beliefs about oneself. The two research directions that constitute the main inspiration for our work - research in psychiatry and computational accounts of decision-making - have been traditionally separate; however recently there has been a growing effort to bridge the distance, and bring computational tools to bear on fundamental research questions in psychiatry. Our aim has been to contribute to this effort, by attempting to formalise and quantify aspects of the interaction between attributions and beliefs about the self.

The first chapter reflects the dual nature of the research that inspired this work: we present the decision-making framework within which attribution and beliefs about the self are conceptualised in our analyses, focusing on the challenges of making decisions and learning from experience in complex environments and on the role that attributions and beliefs about the self have in these processes. We present previous work investigating the way artificial agents, animals and humans address these challenges, and review theoretical accounts pertaining to attributions and beliefs about the self in the psychiatric context, with an emphasis on the research that constituted the background and inspiration for this work. Finally, we present an overview of the computational modelling approach which has been successfully applied in decision-making research, and on which our analyses rely to a large extent.

In the second chapter we present a detailed description of the experiment, highlighting our aims and how they shaped our experimental choices, as well

as the challenges that our choices pose for data analysis, and the suggested improvements for future work.

The third and fourth chapters - the core of this work - are dedicated to the presentation of analyses performed on the two main aspects of the data: subjects' attribution responses and their skill estimates responses, and of the results of these analyses and their interpretation.

The fifth chapter presents analyses of the relationship between behavioural measures and questionnaire scores.

The work concludes with a final chapter, in which we summarise our contribution and present our conclusions, as well as the directions for future work that our results suggest.

Chapter 1

Literature review

The structure of this chapter reflects the dual nature of the research that inspired our work: we begin by presenting the reinforcement learning (RL) framework in which we conceptualise decision making, and provide a simple concrete example of a learning problem in such an environment. We then review literature on balancing exploration and exploitation and assigning credit for experienced rewards and punishments in RL environments and decision-making studies. These are two fundamental problems that RL agents face, and they can be interpreted as concrete and formal instantiations of complex real-world phenomena involving attributions and beliefs about the self. We then review theoretical accounts of attribution that inspired this work, and evidence uncovered by research prompted by these theories. Finally, we present an overview of the computational modelling approach which we use in our subsequent analyses.

1.1 Decision making - a reinforcement learning framework

1.1.1 Environment set-up

Our conceptualization of a decision-making environment is drawn from the reinforcement learning (RL) literature (Sutton and Barto, 2018).

There are a number of elements that are necessary to define a decision-

making environment: first, there is the set of states that it can be in at each moment in time ($\mathscr{S}$), and the set of actions that agents in the environment can perform ($\mathscr{A}$); secondly, there is the mechanism specifying how the environment transitions between states, due to its internal dynamics and/or as a result of actions performed by the agents; finally there are internal mechanisms within the agents which evaluate the desirability -or rewarding value- of the environmental states, or aspects of them (in this formulation, undesirable or punishing states are characterised by negative rewarding values). Agents are assumed to be gathering evidence and adapting their behaviour as a result of interactions with the environment, with the aim of maximising the amount of reward they receive in the environment.

In all but the simplest of cases, irreducible randomness in the world makes both the transitions between states and the effects of actions probabilistic, and in particular also affects the gaining of reward and avoidance of punishment. Therefore the transition mechanism and the gaining of reward are represented as sets of probability distributions: distributions over the next state value, given the current state and current agent actions - $P(\mathfrak{s}_{t+1}|\mathfrak{s}_t, \mathscr{A}_t = \{\text{actions of all agents at time t}\})$; and distributions over the reward, given the current state and actions - $P(r^t|\mathfrak{s}_t, \mathscr{A}_t = \{\text{actions of all agents at time t}\})$. In addition, environments can be more or less stable over time - reward and transition probability distributions can change more or less frequently.

Any of the essential defining components of the environment (including the available actions) could be partially or totally unknown to the agents, imposing therefore the need to learn about it from experience, as well as the need to act in the presence of incomplete information. The fact that environments can change implies that agents cannot exclusively rely on remote past experience as being accurate, but need to be able to detect changes and update their knowledge accordingly. In particular, the range of available actions, and the extent to which they can effect changes in the environment, leading to the gaining of reward or avoidance of punishments, are of fundamental im-

portance to the agent, since knowledge about these aspects of their agency influence the agent's decision-making process. They are, therefore, prime targets of learning. This aspect of learning about the environment is of particular interest in this work, as it is closely linked with attribution.

It is important to note that along with information gained by interaction with the environment, agents might also have prior beliefs or expectations about relevant, but unknown aspects of the environment – this is often the case in real life. Such prior beliefs can encode knowledge that is valid across environments, and thus be beneficial in shaping evidence accumulation and decision-making in new environments. However, they can also hinder learning, if they are inadequate in the new circumstances.

Note that in the discussion above, priors can refer to explicit beliefs, which human subjects can report, but also to implicit expectations that both animals and humans can have, acquired through evolution or during their own lifetime, as a result of prior learning (see our discussion of learned helplessness in section 1.3.2 below).

In the next section we discuss the formalisation of the process of learning from experience in decision-making environments, highlighting some of the challenges that learning agents face in such environments.

1.1.2 A concrete learning problem

The precise formalisation of learning in a decision making environment depends on the particular environment structure, state and action representations and the choice of learning algorithm to be implemented. In this chapter, we seek to illustrate general principles, and so do not present specific details about these elements. However, in order to illustrate how learning from experienced outcomes can be formalised, and how attributions, beliefs and their interactions can be accounted for in the present framework, we will use as a running example a simple concrete learning problem, described below[1].

[1] This example is a tool to illustrate the concepts we are interested in, and their importance; it is not a toy representation of our experiment.

We will use one of the simplest types of decision-making environments, namely one in which there is only one state and two actions, each action a_i providing a reward $r_i > 0$ with a given probability p_i, and no reward otherwise. An agent placed in this environment, and allowed to repeatedly take actions and obtain rewards, will aim to gain as much reward as possible, by choosing appropriately among the two actions to perform on any given trial. This setup is known in RL literature as a "2 armed bandit" problem.

We assume that the agent maintains and updates an internal estimate of the value of each action, and denote by Q_1^t and Q_2^t the estimates of these values at the beginning of any given trial t. We assume that on every trial the agent chooses which action a^t to perform by using a softmax function, according to which the probability of choosing action 1 is

$$p(a^t = 1|Q_1^t, Q_2^t) = \frac{\exp(\beta Q_1^t)}{\exp(\beta Q_1^t) + \exp(\beta Q_2^t)}.$$

According to this action choice mechanism, the action with the highest current estimated value is more likely to be chosen, but occasionally the action estimated to be less valuable can also be taken; the β parameter captures the sensitivity of choices to action values: the higher β is, the more pronounced is the preference for the action currently appearing to be better.

We assume that the agent updates the estimated value of the chosen action after every experienced outcome, by using a simple learning algorithm: it computes the difference between the reward it obtained and its estimate of the value of the action it took, and uses this difference to adjust the estimated value of the action:

$$\delta = r^t - Q_{a^t}^t$$
$$Q_{a^t}^{t+1} = Q_{a^t}^t + \alpha\delta,$$

where a^t and r^t are the action taken ($a^t \in \{1,2\}$) and reward obtained in trial t, and $Q_{a^t}^t$ and $Q_{a^t}^{t+1}$ are the old and updated estimates of the value of action

a^t. The learning rate parameter α controls the weight that new information has, and therefore the speed of learning, higher α meaning higher sensitivity to new information and faster changes in estimated values as a result of experience.

Intuitively, the agent's causal attribution for the experienced outcome is an important determinant of the learning rate. In order to illustrate this, let us assume there is some degree of variability in action execution, and criteria for registering actions as valid: let us assume that the actions are reach movements towards two given targets, and only movements ending within a given area around the targets are recognised as valid. In this case, an action followed by no reward could be due to a probabilistic omission of reward as per the reward schedule of the chosen target (outcome attributed to action choice), but it could also be a result of the movement not reaching the chosen target region (outcome attributed to action execution) (see our discussion of Parvin et al's study (Parvin et al., 2018) in 1.2.2, and section 1.4.2 for further considerations on the distinction between action choice and action execution).

An outcome of no reward conveys different amounts of information about the value of the chosen action, depending on the causal attribution: if it is attributed to the action choice, it is informative about the reward frequency associated with the chosen target; if, however, it is attributed to action execution, it is not informative about the chosen action. We formalise this by using A^t to denote the causal attribution the agent makes on a given trial t, $A^t \in \{\text{choice (c), execution (e)}\}$, and by assuming that the two possible attributions correspond to different learning rates α_c, α_e.

The value update therefore becomes:

$$\delta = r^t - Q^t_{a^t}$$

$$Q^{t+1}_{a^t} = \begin{cases} Q^t_{a^t} + \alpha_c \delta & \text{if } A^t = \text{c} \\ Q^t_{a^t} + \alpha_e \delta & \text{if } A^t = \text{e}. \end{cases}$$

We now specify the way causal attributions are generated, which will

also illustrate the effect that beliefs have on this process. We assume that along with estimates of the values of the two actions, the agent also maintains and updates an estimate of its reach accuracy, or skill, and that it uses this estimate, along with the experienced outcome, when making causal attributions[2]. Specifically, if a reward is obtained, the agent infers that its action was registered as valid, and attributes the outcome to the action choice; if no reward is obtained, the agent bases its attribution on its current belief about its skill, such that the more skilled it believes itself to be, the more likely it is to assume that the reach movement was accurate, and therefore attribute the outcome to the action choice:

$$P(A^t = c) = \begin{cases} 1 & \text{if } r^t \neq 0 \\ \sigma(s^t) & \text{if } r^t = 0 \end{cases},$$

where s^t is the skill estimate at the beginning of trial t and σ is the sigmoid function, $\sigma(x) = \frac{e^x}{e^x+1}$. Note that this is a cartoon example, containing the minimum complexity needed to illustrate relationships between variables of interest; causal attributions in realistic scenarios can be based on several diverse sources of information, and beliefs about the situation, the context, other agents etc.

Finally, we need to specify the update mechanism for the agent's belief about skill. We will assume that this is done in a way similar to the value updates: the agent computes a difference between the experienced and expected accuracy of the reach movement, and uses this error to update its skill value:

[2]Beliefs are usually represented via probability distributions, however in keeping with the point estimate representation of action values which we have used above, we will consider the belief about skill to also be encoded by a point estimate. In this case, the agent's prior experience of agency is only encoded in the initial value of the skill estimate; richer accounts of the agent's belief, such as including the strength of the belief, or indeed representing it as a full probability distribution, can account for more complex effects of prior beliefs, more complex belief dynamics, and richer effects of beliefs on learning, than our current choice allows.

$$\delta_s = \begin{cases} 1 - \sigma(s^t) & \text{if } A^t = \text{c} \\ -\sigma(s^t) & \text{if } A^t = \text{e} \end{cases}$$

$$s^{t+1} = s^t + \alpha_s \delta_s,$$

where α_s is a learning rate for skill. We assume the experienced accuracy is 1 if $A^t = c$, as attributing the outcome to the action choice means the action was performed successfully; conversely the experienced accuracy is 0 if $A^t = e$, as attributing the outcome to execution means the action was not registered due to inaccurate reach movement.

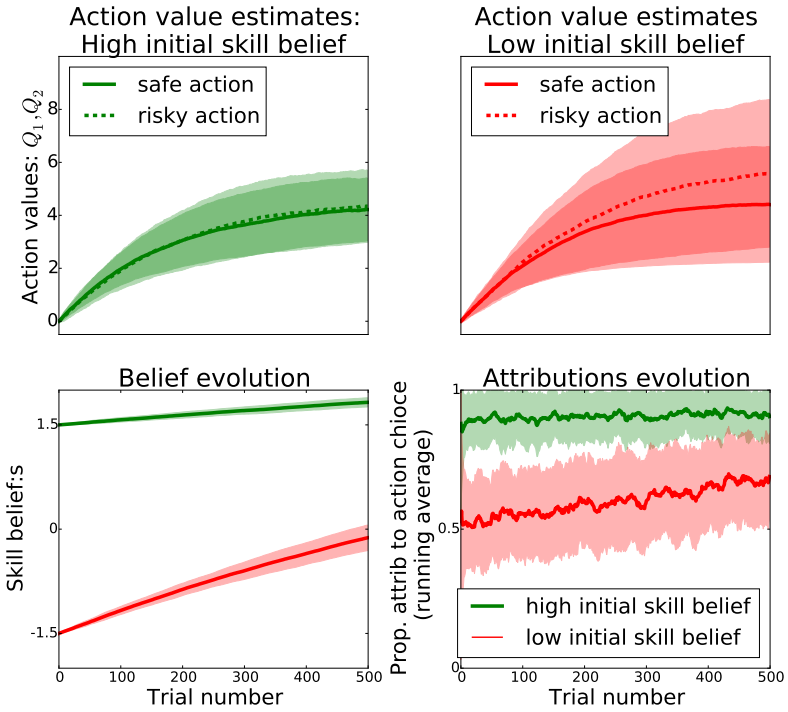


Figure 1.1: Simulation results: mean ± s.d. computed over 100 simulations of the agent described in the text. Task: $r_1 = 10, p_1 = 0.5, r_2 = 15, p_2 = 0.33$; the two actions have equal expected values, but action 1 yields reward more reliably ('safe' action). Agents: two agents were simulated, with identical parameters $\beta = 0.8, \alpha_c = 0.01, \alpha_e = 0, \alpha_s = 0.01$, identical initial estimates for the two action values, $Q_1 = Q_2 = 0$, and different initial estimates for skill:$s = 1.5$ (high initial skill estimate; green) and $s = -1.5$ (low initial skill estimate; red). Top: evolution of action values for the two agents; continuous lines are used for the 'safe' action, dotted lines for the 'risky' one. Bottom: left: evolution of skill estimates; right: evolution of the proportion of outcomes attributed to action choice.

Figure 1.1 shows the evolutions of skill and action value estimates obtained by simulating the agent described above. These simulations illustrate one effect of the difference in beliefs: agents believing they have low skill discount the frequency of non-rewards, as failure to obtain reward is more likely to be attributed to action execution; hence the "risky" action, which yields more reward, but with lower probability, is estimated as being more valuable.

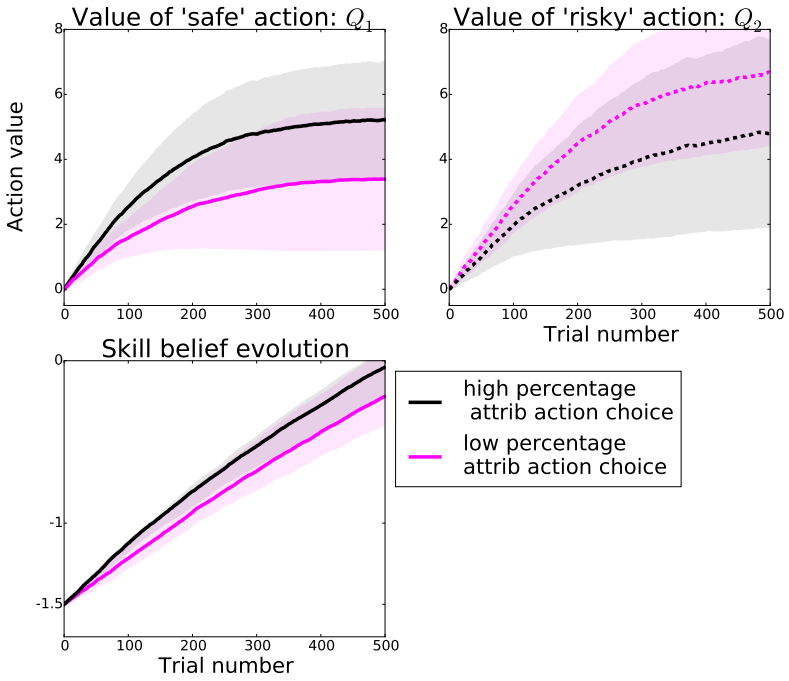


Figure 1.2: Effect of attributions on belief and action value evolutions, agent starting with low initial skill estimate (see caption in figure 1.1 for simulation details). Median split of simulation runs based on percentage of attributions to action choice in the first 125 trials. Black: high percentage of attributions to action choice. Magenta: low percentage of attributions to action choice. Top: evolution of action values; left: 'safe' action; right: 'risky' action. Bottom: evolution of skill belief.

Figure 1.2 illustrates the effect of attributions on beliefs and action values. We split simulation runs for the agent starting with low initial skill values according to the percentage of outcomes attributed to action choice in the first 125 trials (first quarter of whole simulated experiment). More attributions to action choice in the first quarter of trials lead to higher skill estimates. Furthermore, action value estimates also evolved differently, as simulation runs

with more attributions to action choice produced a preference for the "safe" action, unlike runs with less attributions to action choice in the initial quarter of the experiment, which displayed the opposite pattern.

Thus, due to the coupling between attributions and beliefs, early differences between attributions are amplified, leading to different patterns of behaviour.

These simulation results suggest that even in very simple situations, interactions between variables, particularly "loopy" interactions, which involve bidirectional influences between variables, can produce relatively complex behaviour (see 1.3.3 and 1.3.4 for further discussion). Effects similar to the ones illustrated in this cartoon scenario have been experimentally observed in humans (see our discussion of (McDougle et al., 2016, 2019) and (Parvin et al., 2018) in 1.2.2).

The purpose of the above example was to provide, in the simplest setting, a concrete instance of learning in the RL framework introduced previously, as well as to show how some of the effects that beliefs and causal attributions have on learning can be quantified and accounted for in this framework, and illustrate their importance. The example is not meant as a model for our task; we provide a complete account of the precise formalisation of these phenomena in our analyses in the relevant chapters of this thesis (see 3.3 and 4.3). For further discussion on modelling in this framework, see section 1.4 below.

Learning in a decision-making environment can be challenging in many ways. Two of the common problems agents need to solve in decision-making environments, balancing exploration and exploitation and assigning credit for experienced rewards, are intimately related to the evolution of beliefs, and to causal attribution. In the following section we briefly review research into how artificial agents, animals and humans solve these problems in relatively simple settings, before focusing on research on attribution and beliefs about the self.

1.2 Difficulties of learning in RL environments: literature review

1.2.1 Exploration and exploitation

One of the best-known problems that agents need to solve is the "exploration-exploitation dilemma": in an environment that is only partially known to the agent there might well be states and actions as yet unexplored, yielding more reward than the states and actions the agent has already experienced and found to be desirable; furthermore, states and actions that had in the past been undesirable and have thus not been recently visited might have undergone changes making them more rewarding.

The agent therefore has to find a balance between exploiting states and actions which it already knowns to be rewarding, and exploring the environment, or exploring its own repertoire of actions, in order to allow the discovery of better states or actions. Exploration is potentially rewarding, but also risky, as the agent does not know whether better options exist, how much exploration is needed to find them, and what the cost of this exploration will be. Exploiting the known rewarding states and actions provides, instead, a safe, but potentially less valuable course of action. Previous experience and the resulting beliefs and expectations are important in shaping behaviour, particularly as the agent faces this dilemma in the absence of complete information.

In the case of artificial RL agents, there are a number of techniques designed to provide solutions for the exploration-exploitation dilemma (Sutton and Barto, 2018), such as adapting the agent's learning rate, injecting a small amount of randomness in the agent's policy, directly rewarding exploration, adopting optimistic initial guesses for the values of available actions. One solution which efficiently balances exploration and exploitation is Thompson sampling (Thompson, 1933; Russo et al., 2018; Thompson, 1935) - a method which directs exploration towards actions that are likely to be optimal (Russo et al., 2018).

Choice patterns produced by this method are similar to "probability matching" behaviour observed in animals (Herrnstein, 1961; Lau and Glimcher, 2008) and humans (Vulkan, 2000): in a two armed bandit task where the two arms give reward with given probabilities summing to 1, an agent which does probability matching chooses each action with a frequency proportional to its probability of yielding reward; while this behaviour is not optimal in the given task, where the optimal policy would always choose the action most likely to produce reward, it can be advantageous in imperfectly known or changing environments (Sugrue et al., 2004).

Studies with human subjects have shown that subjects' beliefs about the underlying mechanisms of the experiment have an effect on whether subjects probability match or not, people showing less probability matching when they had stronger beliefs in the randomness of the outcomes (Vulkan, 2000; Morse and Runquist, 1960), and when the task was framed as a gambling, versus a skill task (Goodnow, 1955). Research on humans subjects has also uncovered relationships between exploration behaviour and impulsivity (Sadeghiyeh et al., 2020), compulsivity and addiction (Addicott et al., 2014, 2017; Morris et al., 2016).

Disruptions in explorative behaviours have been documented in patients suffering from psychiatric disorders: schizophrenia patients showed less exploration than controls in situations where uncertain options could prove better than the known ones, and the impairment correlated with anhedonia scores (Strauss et al., 2011); patients with depression have been found to explore more than controls in a task involving only gains (Blanco et al., 2013), but also to show less adaptive exploration (Cella et al., 2010; Huys et al., 2012), in tasks involving both gains and losses, hinting at the complexity and subtle dependencies of these effects.

1.2.2 Credit assignment

Another major and multifaceted challenge that agents face in learning in decision-making environments is correctly assigning credit (Minsky, 1961)

for experienced rewards or punishments. There are multiple levels and dimensions along which credit assignment needs to happen.

Time is one of the dimensions, as often rewards or punishments are not the direct consequence of only the most proximal action that an agent took, but are the result of their behaviour over a longer time period; credit for the experienced outcome therefore needs to be divided among the several actions in the chain of behaviour leading to it, which is not a trivial problem to solve. Another dimension refers to the extent to which the agent's actions contributed to producing the outcome, as opposed to it being a result of environment dynamics, or a consequence of irreducible randomness; in environments where multiple agents interact, this problem is further complicated by the possibility of outcomes being the result of another agent's actions, or the result of a combination of actions performed by a group of agents. This aspect of the problem is intimately related to the notions of control and responsibility, and particularly relevant for attribution.

Another, perhaps more subtle dimension of credit assignment, corresponds to the distinction between action identity and action performance and is more obviously relevant for the assignment of blame: when one action fails to elicit reward, is that due to the action identity or to the way the agent executed it? The distinction is often absent from simplified experimental setups involving button presses as actions, but is present in real-life situations, and has been studied in motor learning tasks (Parvin et al., 2018; McDougle et al., 2016, 2019).

Temporal credit assignment: theoretical solutions

In artificial RL agents, there are two kinds of solutions to the temporal credit assignment problem: temporal difference learning (TD) allows agents to learn the values of states and actions, including correctly assigning credit over time, through multiple repetitions of pairings between states and actions; in contrast, eligibility traces allow reward information for each outcome to be propagated back (to varying depths) through the succession of actions and states visited, thus enabling the agent to learn faster (Sutton and Barto, 2018;

Lehmann et al., 2019).

Temporal credit assignment: behaviour

Consistent with the existence of a time-window during which past actions are eligible for reinforcement (Yagishita et al., 2014), research in both animals and humans has found effects of the delay between action and reward on the reinforcement of the action, whereby increasing the delay between reward and action or stimulus preceding it produces diminishment in the reinforcing effect of the reward (Kamin, 1961; Dickinson et al., 1992). Conversely, temporal proximity between an action and a subsequent reward has been found to produce reinforcing effects even in the absence of contingency, a phenomenon known as "spread of effect" (Thorndike, 1933), observed in both animals and humans, leading in some cases to the acquisition of "superstitious behaviours" (Skinner, 1992). Thus in an experiment by Jocham et al (Jocham et al., 2016) with human subjects, subjects' preference for actions performed shortly before the delivery of rewards increased even when actions and rewards were not contingently linked (note that in this case subjects were informed about the contingency structure of the task, and their behaviour showed significant effect of contingency on choices). This coarse form of credit assignment based only on temporal proximity might be constantly in use, but kept in check by more accurate credit assignment mechanisms, and revealed through subtle analyses (Walton et al., 2010), or becoming apparent when these more precise mechanisms are impaired (Devenport, 1979; Kovach et al., 2012; Noonan et al., 2017).

Temporal credit assignment: neural implementation

There is ample evidence that the phasic activity of dopamine neurons in a range of mammalian species presents crucial functional similarities with TD error signals (Schultz et al., 1997; Dayan and Niv, 2008; Glimcher, 2011; Schultz et al., 1993; Takikawa et al., 2004; O'Doherty et al., 2003). A number of candidate mechanisms for neural implementation of eligibility traces (indeed the concept of eligibility traces in RL was inspired by ideas from neuroscience (Klopf, 1972, 1982) in the first place) have been proposed, from spiking activity patterns (Asaad et al., 2017) to synaptic plasticity, notably

the reward-modulated spike-timing-dependent-plasticity (STDP) (Izhikevich, 2007; Reynolds and Wickens, 2002; Pawlak et al., 2010; Pawlak and Kerr, 2008).

Structural credit assignment

Credit assignment among multiple simultaneous cues, or among stimuli or actions varying along multiple dimensions is both particularly difficult and particularly relevant to attribution, since most real world problems in which attribution is relevant involve complex, high dimensional situations, stimuli or actions. In such cases, even when only some of the features are relevant, their identity is unknown, and therein lies the difficulty. RL agents often benefit from a setting of the environment in which state and action representations, hand-crafted by humans based on knowledge of the relevant task-aspects, only incorporate relevant features. However this is not the case for animals and humans, who also need to solve such problems. One relatively simple strategy involves reducing complexity by focusing on a low number of features at a time and testing hypotheses about their relevance, perhaps using prior experience or emotional salience (Heider, 1982; Bentall, 2003; Kelley, 1967) to guide this serial hypothesis testing process. Recent studies investigating this problem quantitatively in simplified settings (Akaishi et al., 2016; Wilson and Niv, 2012) found evidence suggesting that humans employ a serial hypothesis testing strategy, focusing on one hypothesis at time, which is then either confirmed or disproved, in which case attention switches to other hypotheses.

Action identity vs action execution

Evidence for the complexities of credit assignment when a distinction between action identity and action execution is possible is provided by recent research on credit assignment for failures in the context of motor learning (McDougle et al., 2016; Parvin et al., 2018; Mushtaq et al., 2019; McDougle et al., 2019). In a typical two armed bandit context where choices were expressed through button presses and the two arms had equal expected value, subjects displayed a marked preference for the safe option, yielding less reward, but doing so more reliably (Niv et al., 2012; McDougle et al., 2016, 2019); however in a second condition where choice was expressed through a

reach movement towards one of the two targets, therefore allowing variability in movement execution accuracy, subjects preferred to aim for the riskier target, which yielded more reward, but less frequently[3] (the reward probabilities and magnitudes for the two targets were predetermined, and feedback was provided accordingly, irrespective of subjects' real movements; care was taken for the feedback to be credible).

It is important to note that this risk preference was only displayed when subjects were provided veridical visual feedback about the accuracy of their reach movements; in another condition (C3), where choice was also expressed through reaching, but visual feedback was not provided, subjects displayed no preference for either option. This is consistent with a scenario where subjects have optimistic beliefs about their ability to reach a given target, as well as a prior belief that reaching any of the targets is equally difficult; under such assumptions, given information about their movement execution errors, they can try again and do better, hit the more valuable target and get more reward.

Action identity vs action execution: beliefs about agency

Additional evidence for the importance of subjects' belief in their agency is provided by a follow-up study by Parvin et al (Parvin et al., 2018). In this case subjects' choices were always expressed through reach movements, but task instructions were manipulated, such that subjects were told either that hitting the target depended on their reach accuracy, or that it was independent of it. It is important to note that feedback on miss trials was either not provided at all, or provided as a non-informative display of the aiming dot, always placed in the middle between the two targets. Parvin et al found an effect of the agency manipulation, but no effect of the feedback manipulation. Thus, when they were instructed that movement execution was irrelevant, therefore effectively removing movement accuracy as a possible explanation for failure to obtain reward, subjects preferred the safe option. However, this preference was abolished when subjects were instructed that outcomes depended on ac-

[3]Note similarities with our simulations for the bandit example in 1.1.2. Our model could be extended to suit these experiments, by allowing it to account for feedback on reach accuracy and subjects' expectations about the way their ability can evolve.

tion execution; in this case subjects displayed no preference for either option.

Action identity vs action execution: nature of feedback and interaction with beliefs

Given that subjects were instructed that the accuracy of their reach movement determined whether they hit the target or not, a preference for the risky choice might have been expected in this condition. The lack of preference for the risky choice in this case can be attributed to the fact that subjects were not provided with veridical feedback as to the quality of their reach movements; as such this condition is equivalent to the C3 condition in the first experiment, as were the observed effects. This scenario is consistent with evidence from more recent research (Uludag, 2019), investigating the effect of the nature of feedback on learning and decision making in tasks involving motor control to express choices.

These results illustrate the importance of subjects' beliefs about agency during learning and decision making, and provide a quantitative measure of this effect in the particular context of motor learning. They also highlight the complex nature of these beliefs, of their relationship with feedback and of their effects on behaviour: a combination of belief in agency and feedback that they could use to improve their performance was necessary in order for subjects to display perseverance for the risky choice. Furthermore, they provide evidence for the sensitivity of these effects: relatively subtle variations between conditions could lead to qualitatively different behaviour.

As the research summarised above illustrates, credit assignment problems are hard to solve, and they are further complicated by the fact that agents do not have complete information about the environment, the other agents, and even potentially about their own abilities. In such cases agents need to rely heavily on their beliefs and expectations in order to simplify the problem. Therefore beliefs, which are reinforced or changed based on experience, also contribute, via their influence on credit assignment, to the evolution of the agent's policy, and therefore to shaping the agent's experience. Such "loopy" situations - where variables are mutually connected - can support complex dynamics, and also create the potential for the apparition of vicious circles (see

our discussion of learned helplessness in section 1.3.2 below).

1.3 Attribution and self-beliefs

We have provided the description of the general framework in which we consider decision-making, and we have presented challenges that agents face when they need to choose actions and learn from their experience in such environments. We have also succinctly illustrated the importance of expectations or beliefs, and of assigning credit for the experienced rewards or punishments.

We now turn to the concepts of attribution and beliefs about the self, which are central to this work. We begin this section with general considerations about the relevance of attributions and beliefs about the self and about the importance of the relationships between them. We then review the phenomenon of learned helplessness, which was of fundamental importance in the emergence and development of the theories of attribution which have inspired this work: we provide an account of the discovery of learned helplessness, of the mechanisms that have been proposed to account for it, and of the current understanding of it, gained as a result of detailed neurobiological investigations. We continue with a review of the theoretical accounts of attribution which have evolved out of the learned helplessness hypothesis: the revised learned helplessness hypothesis, the learned hopelessness theory, and the attributional self-representation cycle theory. Finally, we review empirical evidence uncovered by research designed to explore and test these hypotheses.

1.3.1 Relevance of attributions and beliefs about the self

As illustrated above for more general aspects of credit assignment, attributions are essential in interpreting experience and making sense of the world and oneself in it; as such, they are crucial determinants of one's expectations about the future, as well as of one's beliefs about the self. In turn, beliefs about one's abilities are likely to influence the evaluation of available actions and their likelihood of success and therefore the decision-making process;

they are also able to influence the post-decision credit-assignment process - the way one makes attributions for one's experience.

Note that while attributions and beliefs about the self are ingredients of decision-making problems, and are relevant as carriers of information, they are also likely to carry important emotional charges, as they directly involve the self. As such, they can be subject not only to constraints related to information processing, but also to constraints of a different nature, related to aspects of (emotional) well-being or motivation. Indeed there is ample research - some of which we discuss below- on biases related to these processes, and the extent to which they can best be interpreted in motivational, rather than strictly information processing terms (Miller and Ross, 1975; Campbell and Sedikides, 1999; Dunning et al., 1995; Zuckerman, 1979). Furthermore, a broad range of external sources of information or decisional factors can be limited to particular circumstances or temporal contexts. In contrast, attributions and beliefs involving the self are more likely to reveal or encapsulate knowledge which is embedded in an individual's identity or self-construct, and thus have a more general impact on behaviour. That is not to say that attributions or beliefs involving the self are necessarily fixed; rather our aim is to highlight that the stakes involved are likely to be higher when the self is concerned than when it is not.

A simple concrete example can be used to clearly illustrate these two processes and their interactions, as well as their emotional and motivational valence: consider a student who finds out their result at a math test – a low mark. If the student has a strong belief in their mathematical ability or general intelligence they might assign this low grade to the test being excessively difficult, to the questions being confusingly expressed, to having had trouble focusing due to tiredness or to some other external or transient potential cause. In contrast, a student who is hesitant about their ability in maths, or who harbours a belief that they are stupid, might interpret the low grade as a result of their poor performance, reflecting their lack of ability or stupidity.

These two alternative scenarios illustrate the effect of beliefs on causal attribution. Consider now the effect of different attributions on beliefs about the self: assigning the low grade to external factors or to tiredness will most likely not affect the subject's estimation of their ability in math, however attributing the low grade to stupidity or lack of ability is more likely to strengthen the student's belief in their inability, and impact their motivation for studying this topic in the future.

This is a clearly a very simplified view of the situation, and these processes can be much more complex and nuanced in practice. Aspects such as whether the cause is transitory or not, the student's belief about whether mathematical ability is something that they can change or not, the importance that math skill has to them personally etc all contribute to the effect of this experience on the student and on their subsequent expectations and behaviour. If, rather than inferring they are stupid, the student believes that ability to solve math problems is not a given trait, but can be acquired, a low grade might be particularly motivating, determining them to work harder and do better in future tests.

This simplified example is meant to illustrate the importance of attributions and beliefs about self for learning and decision-making and the ways in which the two interact. It also shows that these are processes happening constantly in everyday situations, and that even in apparently simple such situations there is potential for a great deal of complexity and variability in the mechanisms of attribution and in their effects.

Both attribution and beliefs about the self are very general concepts, and have been investigated in a huge amount of studies, and in multiple diverse research areas.

Concepts of self are fundamental, but difficult to define and measure, due to the complexity involved. A number of methods have been used to measure aspects of beliefs related to the self, from explicit questionnaires measuring self-esteem, or perceptions and expectations about the self (Rosenberg, 1965;

Higgins, 1987) to more implicit methods such as measuring reaction time for the recall of self-related words (Lyon et al., 1999), or measuring impairment in color-naming self-related words - "the emotional Stroop" effect (Bentall and Thompson, 1990; French et al., 1996).

Attribution-making has also been studied in diverse areas: it is related to general causal inference mechanisms, of which it constitutes a particular case (Heider, 1982); as a key part of the process of forming judgments about others and about their responsibility in different circumstances, it has been a topic of research focused on the assignment of blame (Lagnado and Channon, 2008; Kominsky et al., 2015); to the extent that it is related to the appraisal of important life events, and to the forming and maintenance of expectations about agency and control, it has been investigated in relation with carreer success (Lyons et al., 2020) and health outcomes (Thompson, 1981; Berglund et al., 2014) and constitutes a topic of particular interest in research on psychiatric disorders. This last conceptual framework is the one that we will focus on, due to our interest in the interaction between attributions and beliefs about the self.

One research direction which occupies a central role in the theory of attribution, and has prompted ample empirical research efforts, concerns accounts of the links between depression and attributions and beliefs about the self and the world. In the following sections we review the development of these theoretical accounts, from their start in the learned helplessness phenomenon, through later refinements and expansions. We then review evidence uncovered by the empirical research that these theoretical accounts have inspired.

1.3.2 Learned helplessness

Learned helplessness is a phrase coined by Seligman and Meier (Seligman and Maier, 1967), which refers to the effects that experiencing uncontrollable aversive events has on subsequent avoidance and escape behaviour. Inspired by observations on the effect of strong inescapable stressors on rats (Richter, 1957), as well as by studies documenting effects of Pavlovian fear condition-

ing on subsequent acquisition of escape/avoidance responses (Leaf, 1964), Overmier and Seligman (Overmier and Seligman, 1967) gave a group of dogs held in a harness electric shocks which they could do nothing to prevent or stop; they then placed the same animals in a shuttle box in which they were also administered electric shocks, however these shocks the dogs could escape by jumping in the opposite side of the shuttle box. Overmier and Seligman observed significant differences in escape behaviour between dogs that had been exposed to inescapable shocks previous to the shuttle box testing and dogs which had not. Previously shocked animals were slower to escape, and also presented dramatic changes in behaviour when encountering shocks in the shuttle box: unlike non-shocked animals, which barked and moved agitatedly in the box prior to escaping, the pre-shocked animals soon gave up and passively waited for the shock to finish; in addition, while in non-shocked animals an accidental successful jump over the shuttle box barrier, resulting in shock termination, was reliably followed by faster subsequent escape responses, in pre-shocked animals successful accidental escapes did nor reliably predict future escape responses.

In a series of studies aimed at uncovering the mechanisms responsible for these differences, Seligman and Meier introduced an additional control group - animals given escapable shocks previous to shuttle box testing; thus the refined version of the experiment involved three groups of animals: one group given no shock prior to shuttle box testing, one group given shocks which they could end by pressing a lever, and a yoked group, experiencing exactly the same shocks as the previous group, but for which shock termination was independent of their own behaviour. Seligman and Maier (Seligman and Maier, 1967) observed that the inescapably shocked animals were later impaired in avoiding and escaping shocks in the shuttle box, unlike the other two groups. When experiencing shocks in the shuttle box, animals previously given no shock and animals previously given escapable shocks moved frantically and, after stumbling upon the correct escape response, quickly learned

from it. In contrast, animals given inescapable shock quickly gave up moving in the shuttle box, and even if they occasionally produced the correct response, they did not learn from these instances of success, but continued to passively accept shocks. These results were interpreted as proof that it was the inescapable nature of the shocks, rather than the shocks themselves, that produced the dramatic behavioural changes. A large number of studies by Seligman, Meier and Miller as well as by other groups showed the effect to be robust and reproducible across different species (Maier and Seligman, 1976) - rats (Seligman et al., 1975), fish (Padilla et al., 1970), cats (Thomas and Baiter, 1974), humans (Hiroto, 1974; Miller and Ross, 1975; Miller and Seligman, 1976) - in experiments involving a range of uncontrollable stressors(swim, restraint, electrical shocks, noise, unsolvable anagrams) and test behaviours (anagram solving, competition for food, dominant behaviour etc).

Seligman and Meier (Maier and Seligman, 1976) proposed the learned helplessness theory to account for these phenomena: in contrast with associative theories of learning, they postulated that animals are able to detect and learn not only contingency between responses and outcomes, but also the lack of it - uncontrollability - whereby $p(\text{outcome}|\text{response}) = p(\text{outcome}|\text{absence of response})$. They further postulated that the cognitive and motivational effects of this learning were responsible for the impairments observed in these animals' escape response: having detected that it has no control over the outcomes, the animal expects the lack of control to be present in the future; this reduces the motivation to initiate escape responses, since the animal expects responses to be useless; in addition, the active learning about uncontrollability interferes with subsequent learning, and animals who have learned helplessness are unable to learn from occasional successful escape responses in the way that normal animals do. Drawing on similarities between the cognitive, motivational and emotional effects of learned helplessness and symptoms of depression, the learned helplessness theory of depression was proposed (Seligman, 1972; Miller and Seligman, 1975; Abramson

et al., 1978), which gave rise to a research direction we review in the following section.

A number of alternative interpretations for the learned helplessness phenomena were proposed: adaptation or sensitisation to shocks, movement impairment due to depletion of neurotransmitters during initial uncontrollable stress, accidental learning of associations between some responses and shock termination during inescapable stress, responses which then contradict the ones necessary for escape in the shuttle box. However, a number of targeted studies seemed to conclusively prove that the explanation based on learning of uncontrollability better accounted for the facts (see Maier and Seligman, 1976, for review) for review.

However later investigations into the neural basis of learned helplessness, which uncovered some of the mechanisms underlying it, showed that Seligman and Meier's initial explanations were only partially correct. The picture that emerged as a result of these later studies suggests that while it is true that the controllability dimension is essential, it is not uncontrollability that is learned, but control. According to this latter view, prolonged aversive experiences lead by default to passivity, as well as increased vulnerability to stress in the future; detecting control over aversive stimuli counteracts this default reaction, protecting from impairments in escape learning and promoting expectations of control for future experiences. The neural implementation of these processes involve the dorsal raphe nucleus (DRN) and its connections with the frontal cortex. Intense aversive stimuli produce an increase in the activity of serotonin neurons in the DRN, producing two effects. The immediate effect is inhibition of the periacqueductal gray (PAG), a structure responsible for generating the fight and flight response, leading to immediate passivity. A second effects involves a longer timescale, and manifests as a persistent generalisation of passivity across contexts; this is produced due to sensitization of the DRN to stress, which facilitates the heightened inhibition response to later stressful events. However detection of control by

a circuit involving the ventro-medial prefrontal cortex and the dorso-medial striatum inhibits the DRN activity, thus preventing both immediate passivity and sensitisation to later stress. In addition, detection of control over aversive stimulation strengthens the vmPFC to DRN pathway, such that it becomes activated by aversive stimuli alone, producing the immunisation seen in animals previously exposed to controllable shock (see Maier and Seligman, 2016, for a detailed review).

A connection between learned helplessness and depression remains relevant, despite the fact that the initially postulated connection between them needs to be reappraised in view of the current understanding of the mechanisms involved in learned helplessness. Furthermore, our current and future understanding of these mechanisms could suggest treatment approaches. We review the evolution of the learned helplessness theory in depression research in the following section.

1.3.3 Attributional patterns and inferences about the self: theoretical accounts

Learned helplessness theory of depression

Similarities between learned helplessness in animals and depression in humans lead Seligman (Seligman, 1972) to hypothesize a link between the two phenomena. The resulting learned helplessness theory of depression postulated as a central cause of depression the cognitive (expectation of lack of control in the future), motivational (impairment in action initiation, due to expectation of failure) and emotional (heightened stress, depressed mood) effects of the belief that one is unable to control (important) outcomes (Maier and Seligman, 1976). There followed a series of studies (Miller and Seligman, 1973, 1975, 1976) aiming to establish the validity of this hypothesis, by comparing the effects of learned helplessness and depression in humans (though note that depressed subjects were defined as such based on their BDI scores (Beck et al., 1961), rather than through a clinical diagnostic). The results of these studies were to some extent consistent with the theory's predictions, but

also hinted at more complex phenomena, which the theory did not account for (Miller and Seligman, 1973, 1976).

Reformulated learned helplessness theory of depression

Abramson et al (Abramson et al., 1978) proposed a refinement of the theory, featuring attribution as a key ingredient that mediates the effect of lack of control on later behaviour. According to this reformulation, once failure to exert control over aversive outcomes has been detected, it is attributed to a cause, which can be internal to the subject or external, stable or transitory, global or specific. It is the nature of this attribution that determines the extent of expectations of helplessness in the future, with attribution globality determining the generality of helplessness deficits, attribution stability determining the chronicity of helplessness deficits, and attribution internality determining the effect on self-esteem. Abramson et al further postulated that people have stable patterns of attribution-making - "attributional styles", and that people with an internal, global and stable attributional style for negative events would be most vulnerable to depression.

Hopelessness theory of depression

The theory was further revised by Abramson, Metalsky and Alloy (Abramson et al., 1989), who postulated the existence of a subtype of depression - hopelessness depression - and articulated a causal chain of events producing it. According to the theory, hopelessness - defined as an expectation that negative outcomes will occur, while positive ones will not, and that there is nothing that one ca do to change this- is a sufficient proximal cause of hopelessness depression symptoms. The causal chain leading to hopelessness starts with negative life events, or the absence of positive ones, which trigger inferences about their causes and their consequences, and inferences about the self; these inferences, potentially modulated by additional factors, ranging from genetic factors to the presence or absence of a social support network, can produce hopelessness.

In keeping with the previous version of the theory, inferring stable and global causes for important negative events increases the likelihood of hopelessness. Unlike the previous version of the theory, inferences about the

consequences of negative events are postulated to influence the likelihood of hopelessness depression independent of causal inferences, but in parallel ways: inferring that an important negative event will have general consequences unlikely to be averted increases the likelihood of people becoming hopeless. Finally, the revised version of the theory emphasises the importance of inferences about the self as a distinct factor, albeit one potentially related to causal attributions: inferences about one's worth, abilities, desirability etc increase the likelihood of hopelessness to the extent that the individual infers that they have negative characteristics which are important, unlikely to change, and affecting many areas of their life. Individual differences in cognitive styles - encompassing differences in the tendency to make stable and global attributions for negative events and differences in the tendency to infer bad things about oneself- contribute, along with situational information, expectations, motivation, attention etc to inferences about causes and the self, and thus act as distal causes of hopelessness.

Hopelessness theory of depression: diathesis-stress model

One of the key aspects of the theory is that it postulates a diathesis-stress model, according to which it is the conjunction of negative life events and predispositions for negative inferences which produces hopelessness, rather than any of these factors in isolation. As a corollary, the theory postulates a titration model of this interaction: people who are very prone to making negative inferences might become hopeless as a result of relatively benign negative life events, whereas people who are less prone to making negative inferences only become vulnerable to hopelessness when experiencing severe negative life events.

By laying out a causal chain leading to hopelessness and hopelessness depression, the theory also indicated protective factors and interventions which could contribute to prevention and improvement of symptoms, such as protective cognitive styles and attitudes towards the self, and interventions aimed at changing disfunctional attitudes and depressogenic attributional styles. We review findings consistent with and supporting of the theory

in the next section.

The attribution-self-representation cycle theory

The theories we have just reviewed assume stable mechanisms are involved in generating causal attributions and inferences about the self and propose a "feed-forward" causal chain linking negative events to symptoms of depression. Bentall et al (Bentall et al., 2001; Bentall, 2003) proposed a theoretical account including interactions between attributions and beliefs about the self, and emphasising the dynamic nature of the processes involved - the attribution-self-representation cycle theory. Bentall et al (Bentall et al., 2001) postulate that rather than having a trait-like attributional style, individuals use current beliefs about the self, or readily available stored knowledge about the self, along with attributional signposts in situational information (Kelley, 1967) when making causal attributions. Bentall et al propose that the process of attribution formation involves a search for explanations that starts with explanations involving the self and terminates when a suitable explanation is found; on the other hand, once an attribution is formed, it primes representations of the self that are consistent with it. Thus, along with effects of attributions on beliefs about the self, Bentall et al recognise the possibility of effects in the opposite direction, leading to a system with fluctuating components and potentially complex effects of interactions between them - the attribution-self-representation cycle (Bentall, 2003). The system can be influenced by relatively stable factors, such as individual differences in stored knowledge about the self, motivational biases, tendency to attend to specific types of information or ability to understand others, as well as variable factors determining the relative availability of information in different circumstances.

Within this framework, self-serving or self-enhancing attributional biases that healthy people display function as homeostatic mechanisms for the maintenance of beliefs about the self within healthy parameters; the absence or disruption of these homeostatic mechanisms leaves people vulnerable to aberrant protective mechanisms or vicious cycles (where negative internal attributions lead to a worsening of self-beliefs, leading to further negative at-

tributions), producing and or maintaining depressive or paranoid symptoms. The dynamic nature of both attributions and self-beliefs, and the fact that they exert reciprocal influences on each other can be expected to produce complex patterns of relationships between them; the inherent difficulty of predicting such relationships can account for some of the inconclusive or contradictory findings of studies aimed at testing previous theories, which suggested poor predictive power (see Robins and Hayes, 1995; Liu et al., 2015, for reviews).

1.3.4 Attributional patterns and inferences about the self: empirical evidence

Investigations into learned helplessness and the emergence of various theories postulating links between this phenomenon and depression prompted a large amount of research in the area, producing a complex pattern of findings, not entirely explained by either theory, which highlights the importance of longitudinal studies, and the need for a complex and nuanced approach. We review some of the empirical results in this section.

Depressive realism

Early investigations into potential differences between the way depressed and non-depressed subjects perceive the amount of control they have over outcomes (a key determinant of learned helplessness phenomenon, as well as an important factor contributing to the formation of causal attributions) uncovered a surprising result: according to learned helplessness theory, depressed subjects were expected to underestimate the extent to which they were in control of outcomes, unlike controls. However this is not what Alloy and Abramson (Alloy and Abramson, 1979) found in their study investigating the perception of contingency between outcomes and behaviour in normal and depressed subjects. Instead, depressed subjects turned out to be accurate in their perception of contingency, unlike normal subjects, who showed a doubly inaccurate pattern of contingency perception, overestimating contingency when appetitive outcomes were involved, and underestimating it for aversive outcomes. These results have been replicated (Martin et al., 1984; Vázquez, 1987) and they are consistent with amply documented self-serving biases in

normal subjects (see Campbell and Sedikides, 1999, for a review).

Depressive realism challenged

However other studies have failed to find associations between depressive symptoms and higher accuracy, or have found only partially consistent results (Alloy et al., 1981; Alloy and Abramson, 1982; Benassi and Mahler, 1985; Presson and Benassi, 2003) (and Allan et al., 2007; Moore and Fresco, 2012, for reviews). Note that most of the above mentioned studies did not use clinically depressed subjects, but subjects labeled as such based on BDI indexes; studies involving depressed patients challenged the depressive realism view, providing evidence that depressed patients have negative biases in their judgements which are not incompatible with them appearing more accurate in some cases (Carson et al., 2010; Kaney and Bentall, 1992), and that they share with non-depressed subjects the illusion of control bias - a tendency to estimate they have some degree of control when none is present (Venkatesh et al., 2018; Carson et al., 2010; Moore and Fresco, 2012; Vázquez, 1987; Presson and Benassi, 2003; Kaney and Bentall, 1992).

Depression and negative biases

A range of studies found associations between depression and negative biases, with people presenting depressive symptoms also showing more pessimism, lower expectations of competency (Golin et al., 1977; Miller and Seligman, 1973; Garber and Hollon, 1980), more negative evaluation of the self and negative self-reinforcement (Rozensky et al., 1977; Lobitz and Post, 1979; Gotlib, 1981; Roth and Rehm, 1980), and higher recall and endorsement of negative self-descriptive words (Lyon et al., 1999; Gotlib and McCann, 1984) than healthy subjects. Depressive symptoms have also been repeatedly found to be associated with a negative attributional style (Peterson and Seligman, 1984; Lyon et al., 1999) involving internal, stable and global attributions for negative events (see Sweeney et al., 1986; Robins, 1988; Robins and Hayes, 1995, for reviews). Note however that results of studies with clinically depressed subjects and those with subjects showing mild to moderate depressive symptoms were not always consistent, and a number of studies have failed to find these effects (see Coyne and Gotlib, 1983, for review).

More generally, associations between depressive symptoms and a tendency to make negative inferences has been observed both in clinically depressed samples and in subjects with mild to moderate depression symptoms (see Liu et al., 2015, for a review).

Diathesis-stress model

One of the key aspects of the hopelessness theory of depression is the diathesis-stress nature of the vulnerability to hopelessness, according to which it is the conjunction of negative life events and negative attributional style that increases the risk of depression, rather than each of these two factors on its own. This aspect of the theory has been tested in numerous studies, both as a risk factor, when negative life events and negative inference styles are concerned (Metalsky et al., 1987; Alloy et al., 1999), and as a protective one, when positive life events and a positive inferential or attributional style are involved (Needles and Abramson, 1990; Kleiman et al., 2013; Haeffel and Vargas, 2011; Johnson et al., 2017), and generally found to be valid (see (see Abramson et al., 1989; Robins and Hayes, 1995; Liu et al., 2015; Johnson et al., 2017, for reviews).

Dynamics of attributions

Much of the research we reviewed so far assumed attributional styles to be stable, trait-like characteristics. However more recent research investigating the effect of experience manipulation on subsequent attributions challenged this view, producing evidence that this phenomenon is more dynamic than previously considered (Forgas et al., 1990; Bentall and Kaney, 2005). In a series of experiments in which subjects' mood was manipulated (either through false feedback in an experimental task or through exposure to emotionally charged short films) Forgas et al (Forgas et al., 1990) measured attributions made by healthy subjects either for hypothetical situations or for their own real exam results and found that subjects given a positive mood made more internal and stable attributions for positive outcomes, and less internal and stable attributions for negative outcomes, than subjects given negative mood. In a study involving clinical populations, Bentall and Kaney (Bentall and Kaney, 2005) administered the expanded ASQ questionnaire to de-

pressed, paranoid and normal subjects and asked them to judge contingency between their actions and outcomes in a computer-based task before and after exposing them to a mild failure experience (an anagram solving task which included unsolvable items); negative internality scores increased after the failure experience for both groups of patients, although not for the healthy controls, who might be less vulnerable to the effects of such mild failure experiences. These studies indicate that attribution-making tendencies can vary rapidly, and that factors such as mood and achievement can exert an effect on them.

Dynamics of beliefs about the self

Beliefs about the self - and their fluctuations - might represent another important source of dynamics in attribution-making. Research involving relationships between attributions and beliefs about the self has found that measures of self-esteem can account for variation in attributional styles in both the general population and in psychiatric patients (Tennen et al., 1987; Tennen and Herzberger, 1987; Romney, 1994), and that situations involving potential threats to self-esteem promoted self-serving attributions (Dunning et al., 1995). These results highlight the need for more precise investigations specifically designed to uncover the dynamics of attributions and beliefs about the self, allowing for quantitative testing of theoretical accounts based on time-varying interactions between these variables, such as the attribution-self-representation cycle (Bentall, 2003; Bentall et al., 2001; Bentall and Kaney, 2005). This need has inspired the work presented in this thesis, our aim being to draw upon precise quantitative measuring techniques typically used in reinforcement learning and decision making research in order to develop a task that would allow such precision to be used in the attribution research domain.

Having reviewed the theoretical accounts of attribution developed based on the learned helplessness hypothesis, along with empirical evidence uncovered by research inspired by these hypotheses, we conclude this chapter with an overview of the technical tools we have employed in this work. Most of the

research we have reviewed so far utilised statistical analyses using averages, rather than trial by trial data. Our purpose in the experiment described in this thesis was to provide a task in which the attributions and self beliefs variables could be measured repeatedly, allowing for investigations of their dynamics by use of more recent trial by trial modelling techniques. We provide a discussion of this approach in the next section.

1.4 Computational modelling - theoretical aspects

To a large extent, the research on attributions and beliefs about the self which we have reviewed above relied, from a technical point of view, on classical statistical techniques such as analysis of variance (ANOVA) and hypothesis testing. We have also consistently employed these techniques in our analyses, as described in later chapters. One of the aims of the work presented in this thesis has been to complement this classical approach by also including in our analyses the powerful and flexible tools of computational modelling, a framework which is particularly well suited to quantify and reveal aspects of the processes underlying phenomena of interest.

Computational modelling has been very popular in research on reward learning and decision making (see Dayan and Abbott, 2001; Dayan and Niv, 2008; Dayan and Daw, 2008; Daw and Doya, 2006; Rushworth et al., 2011; Behrens et al., 2009, for reviews), including work that we reviewed above in section 1.1.2, and has recently been increasingly viewed as a useful framework for psychiatry research (Montague et al., 2012; Huys et al., 2015; Adams et al., 2016; Hauser et al., 2019). It constitutes the theoretical foundation upon which our model-dependent analyses are based. In this section we present an overview of the probabilistic computational modelling framework, highlight the two core components of this approach - parameter estimation and model comparison, and briefly discuss modelling decisions involved in applying the approach at the population level. Detailed accounts of implementation in our

work are presented in the subsequent relevant chapters.

In computational modelling approaches, hypotheses about the mechanisms producing the observed data are encoded into precise mathematical formulations involving observable variables, or variables set by the experimenter, and unknown or unobservable variables. Given a setting of all variables involved, the model can be used to generate predictions, which can then be compared with the real, observed data, and discrepancies between prediction and reality can be measured. Typically, at least some of the model parameters represent unobservable variables of particular interest, which are to be inferred from data: assuming the observed data has been generated according to a given model, this inference corresponds to finding the setting of the corresponding parameters that minimises the discrepancy between predicted and real data. The approach can also be used to answer questions of a different nature: the extent to which different models account for the observed data can be used to evaluate the corresponding hypotheses that the models encode. These two kinds of problems are known as parameter estimation and model comparison respectively.

One key advantage of using generative models compared to classical statistical tools is that they can encode theories accounting for the dynamics of the variables of interest at the trial by trial level. This feature is particularly relevant in the case of learning theories which explain how feedback is related to trial by trial changes in behaviour or neural activity, and the increasing use of such models has enriched our understanding of trial by trial dynamics.

1.4.1 Specifying a probabilistic model

The computational models we use in this work are probabilistic models (MacKay and Mac Kay, 2003; Bishop, 2006; Koller and Friedman, 2009): while some of the relationships between variables in these models can be deterministic, they also allow for the presence of noise, represented through probability distributions, and accounting for inherent uncertainty and or for the effects of nuisance variables.

Noise terms can be included in the relationships between unobserved variables, as well as in the relationships between these and observable data. In addition, these models include uncertainty over parameters (as well as parameter variability between subjects, as discussed below), also encoded as probability distributions. A probabilistic model is therefore formally specified by two such distributions: the likelihood function $p(D|\theta,M)$, which encodes the probability of data D given a setting of the model parameters θ, and the prior distribution over parameter values under the model, $p(\theta|M)$[4].

To illustrate this[5], let us return to the 2-armed bandit problem that we discussed previously from the agent's perspective (see section 1.1.2), and present it from the point of view of an experimenter using a computational approach to analyse the behaviour of a subject. Let us assume the experimenter postulates the subject uses a simple, deterministic learning model, $Q_{a^t}^{new} = Q_{a^t}^{old} + \alpha(r^t - Q_{a^t}^{old})$, with one learning rate parameter α. In oder to generate predictions, the experimenter's model needs to also account for the link between action value estimates and the choice of action on every trial. This link we assume to be a probabilistic one, namely the softmax function introduced before, $p(a^t = 1|Q_1^t, Q_2^t) = \frac{\exp(\beta Q_1^t)}{\exp(\beta Q_1^t)+\exp(\beta Q_2^t)}$, where $p(a^t = 1|Q_1^t, Q_2^t)$ is the probability of choosing action 1 at trial t, given the current estimates of the values of the two actions; this choice introduces the additional parameter β, modelling the sensitivity of choices to action values [6]. Considering the data to be the series of actions the subject chose on every trial, $a^1, a^2, ...a^N$, and assuming that choices are independent given model

[4]Note that here we use the most abstract representation of these distributions, our purpose being to illustrate the theoretical foundations; in practice, these distributions can be arbitrarily complex objects, depending on the specific details of the modelling problem.

[5]We use this example here merely for illustration purposes; the models we use in chapter 3, which are built on a similar structure, are described in detail in the appendix J.

[6]Note that in this case the experimenter's modelling choices match the action selection mechanism that the agent uses, but not the learning model.

parameters, the term $p(D|\theta,M)$ becomes

$$p(a_1,...a_N|\alpha,\beta) = \Pi_t p(a^t|Q_1^t,Q_2^t)$$
$$= \Pi_t \frac{\exp(\beta Q_{a_t}^t)}{\exp(\beta Q_1^t)+\exp(\beta Q_2^t)},$$

where we use $Q_{a^t}^t$ to denote the current estimated value of the action took at trial t, a_t.

1.4.2 Parameter estimation

Parameter inference is most often conceived as an optimisation problem which consists in finding θ such that $p(D|\theta,M)$ is maximised - the maximum likelihood estimator [7]. Finding the maximum likelihood estimator is equivalent to finding θ that minimises the negative log-likelihood, $-\log(p(D|\theta,M))$; this formulation is preferred due to its advantages for numerical optimisation procedures, which are generally used to implement parameter estimation.

In the 2-armed bandit example above, the corresponding optimisation objective is $-\log(\Pi_t p(a^t|Q_1^t,Q_2^t)) = -\Sigma_t \log(\frac{\exp(\beta Q_{a_t}^t)}{\exp(\beta Q_1^t)+\exp(\beta Q_2^t)})$, and solving this optimisation problem would produce estimates of α and β for the given subject.

Our purpose being to illustrate this process in the simplest possible context, we have so far discussed parameter estimation as applied to data from one subject. However real data typically involves not one, but a number of subjects: questions of interest, such as determining whether attributions have an effect on subsequent beliefs about self, or comparing this effect in healthy

[7] An optimisation problem more consistent with a Bayesian approach is maximising $p(\theta|D,M)$, the result of which is known as the maximum a posteriori estimator. The two optimisation objectives are linked through Bayes' rule $p(\theta|D,M) = \frac{p(D|\theta,M)p(\theta|M)}{p(D|M)}$, therefore the distinction between the two optimisation problems is relevant only if the model includes a non-uniform prior distribution over the parameters. Such distributions can be used to encode parameter constraints, or assumptions about likely values, based on prior knowledge. A fully Bayesian approach involves manipulations of the posterior probability distribution, $p(\theta|D,M)$, rather than the use of point estimates of parameters. However for the benefit of clarity and brevity we use point estimates in this discussion, and we assume uniform prior distributions, allowing us to focus on $p(D|\theta,M)$.

and depressed people, are generally not questions about one individual; rather they are naturally formulated as questions about one or more populations. There are several approaches for modelling data from multiple subjects.

One approach consists in fitting each subject individually, by solving the corresponding optimisation problem, which results in a pair of α, β values for each subject, and then applying classical statistical methods to the resulting samples for α and β.

To illustrate the process in our 2-arm bandit problem, let us assume the experimenter asks subjects to report, on every trial, whether they believe their reach movement to have landed in the target region or not. And let us further assume the experimenter postulates a slightly more complex learning model than the one discussed above, namely one which has two learning rate parameters, α_{inacc} - corresponding to trials in which subjects judge their performance of the reach movement to have been inaccurate, and α_{acc}, corresponding to trials where subjects judge their movement to have been accurate. Having obtained parameter estimates for all subjects, the experimenter can test whether the two learning rates are different in the subject population, or whether the α_{inacc} learning rate is different from 0. (Note that we use this example as a very simple and concrete illustration of the way in which model parameters can be used to address questions of interest; the process of building computational models of behaviour in cognitive tasks is generally a more complex endeavour.)

We have used this approach repeatedly throughout this work, as detailed in the following chapters, where we present our analyses and results. Wether subjects use different learning rates for accurate and inaccurate reach trials is a question that can also be framed in model comparison terms, which will be the topic of the next section.

A different, and, from a model fitting perspective, a slightly more sophisticated, method of dealing with population data consists in explicitly accounting for variability between subjects in the generative model. This is

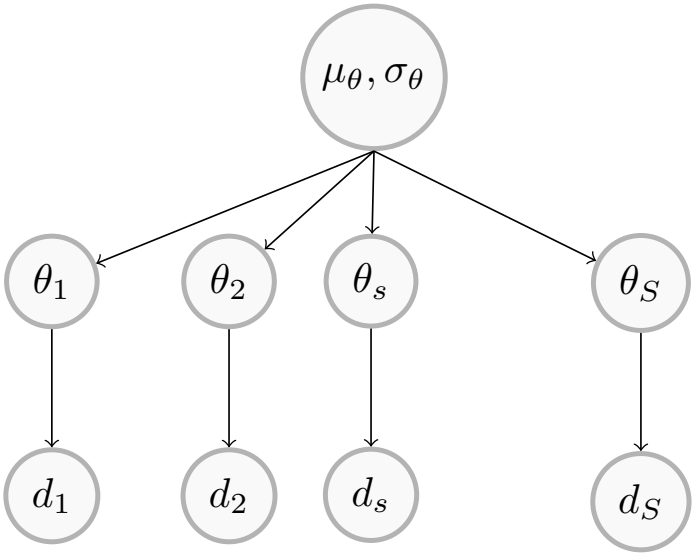


Figure 1.3: Graphical representation of an abstract hierarchical population model: $\mu_\theta, \sigma_\theta$ represent population parameters; $\theta_1, \theta_2, ...\theta_S$ represent individual parameters for S subjects; $d_1, d_2, ...d_S$ represent data for subjects 1, 2, .. and S respectively; arrows represent probability distributions encoded by the model: individual parameters are assumed to be drawn from a Gaussian distribution specified by population parameters, $\theta_s \sim \mathcal{N}(\mu_\theta, \sigma_\theta)$; the link between θ_s and d_s is given by the likelihood in the model, $p(d_s|\theta_s)$.

achieved by augmenting the model with a set of population-level parameters, and with the assumption that individual parameters are drawn from a probability distribution - often assumed to be Gaussian- characterised by these (unknown) population-level parameters. Thus in addition to individual parameters and the likelihood of individual data given individual parameters, the model will also include a term defining the likelihood of individual parameters given population ones (see figure 1.3 for a graphical representation of this more complex model).

In our 2-armed bandit example, this would correspond to the addition of population level parameters $\mu_{\alpha_{inacc}}, \mu_{\alpha_{acc}}, \mu_\beta, \sigma_{\alpha_{inacc}}, \sigma_{\alpha_{acc}}, \sigma_\beta$, representing the mean and standard deviation of the Gaussian distributions of the three parameters at the population level, such that each individual subject parameter is assumed to be drawn from the corresponding distribution (e.g. $\beta_i \sim \mathcal{N}(\mu_\beta, \sigma_\beta)$). In this case, the parameter estimation step would corre-

spond to finding population-level parameters that maximise the likelihood of the data. Testing wether α_{inacc} is different from 0, or wether the two learning rate parameters are different from each other, would be performed on the population distributions defined by the estimated population level parameters.

While the process is theoretically straightforward, in practice it generally involves complications due to the need to integrate out individual-level parameters in order to compute the likelihood function. The corresponding integral is often intractable, as is the case in the example used above, where the population distributions are Gaussian, but the individual likelihood functions are not. In these cases, sampling or analytical approximations are needed (MacKay and Mac Kay, 2003; Bishop, 2006; Gelman et al., 2013). Our aim here being to present the general theoretical aspects of our modelling approach, we do not discuss these issues in further detail at this point. We have used this hierarchical modelling approach in some of the analyses presented in this work, and we provided a more detailed discussion in the relevant sections (see 4.3).

1.4.3 Model comparison

As mentioned above, determining if subjects use different learning rates for accurate and inaccurate reach trials is a question that can also be framed in model comparison terms, an issue to which we now turn.

Model comparison refers to evaluating the extent to which different models capture the mechanisms producing the observed data. It is a fundamental aspect of computational modelling, as it underlies its two key goals: evaluating alternative hypotheses about the phenomena of interest - which translates directly into comparison of the corresponding models, and inferring hidden variables- inferences about parameters are only informative to the extent that the chosen model represents a good approximation of the real data generating process.

Note that a good model is one that captures relevant aspects of the underlying data generating process, rather than the observed data itself: simply

using the quality of model fit to the data as the comparison variable is not a suitable approach. This can be easily illustrated by considering two nested models, such as the two models we described for the 2-armed bandit task, one including only one learning rate parameter, the other including two. As the second model is an expansion of the first one, it will inevitably fit any given data at least as well as the first one, because optimisation is performed over a larger set of parameters. As this is merely an automatic result of the way the comparison is framed, it does not constitute evidence that subjects use different learning rates for the two different types of trials.

One way to perform a fair and informative comparison of the two models consists in comparing their prediction accuracies on new data from the same source. This illustrates a general, straightforward but computationally expensive, approach to model comparison: separating the available data into a training set used for model fitting, and a held-out set, to be used for computing the model accuracy. Comparison between alternative models can then be performed based on their accuracy on held-out data.

Alternatively, in the probabilistic framework, model comparison is performed by computing, for all the different candidate models, a quantity known as the model evidence; the model with the highest model evidence is preferred. Model evidence is defined as the likelihood of the data given the model, and it is computed by integrating over the model parameters according to Bayes' rule: $p(D|M) = \int p(D|\theta,M)p(\theta|M)d\theta$, where $p(D|\theta,M)$ is the likelihood of the data given parameters, and $p(\theta|M)$ is the prior likelihood of the parameters included in the model.

Comparison via the model evidence constitutes an automatic implementation of Occam's razor, preferring a model that is "just right" for the data: for a model that is too simple to produce the observed data, the likelihood term is low for any setting of the parameters; for a model that is too complex, and could produce the observed data, along with a large number of other potential datasets, prior probability density is spread out over many possible settings

of the parameters, out of which only few correspond to accurate predictions on the observed data; a preferred model is one which lies in between these extremes, being just flexible enough to capture the data (see figure 1.4 for a cartoon illustration).

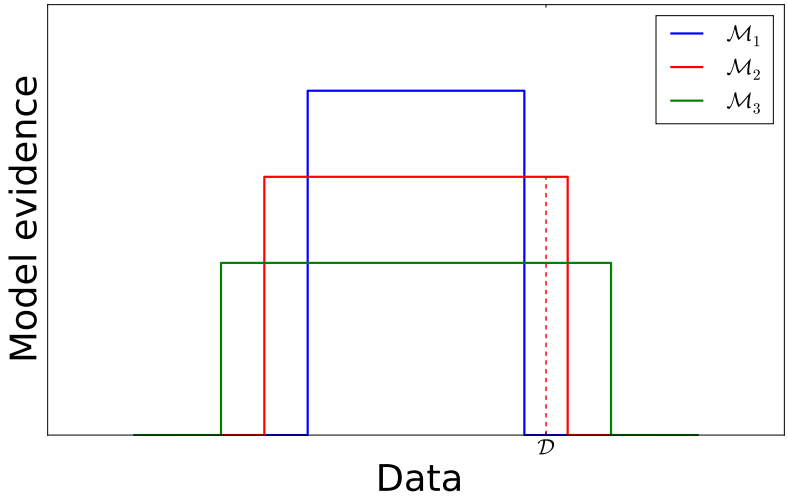


Figure 1.4: Model evidence implementing Occam's razor: for dataset $\mathscr{D}$ model evidence favours model $\mathscr{M}_2$, which is complex enough to fit the data, and not spreading its probability distribution over too large a family of datasets; model $\mathscr{M}_1$ assigns higher probability to the datasets that it can fit, but $\mathscr{D}$ is not among them; model $\mathscr{M}_3$ can fit a larger family of datasets, including $\mathscr{D}$, but assings smaller likelihood to each of these than model $\mathscr{M}_2$ does.

In practice, integrals involved in computing the model evidence are often intractable, and approximation is needed (MacKay and Mac Kay, 2003; Bishop, 2006; Gelman et al., 2013). One of the most frequently used solutions is the use of theoretically derived approximations of the model evidence (Schwartz, 1978; Akaike, 1974; Watanabe, 2013), such as the Bayesian Information Criterion (Schwartz, 1978), $BIC = k\ln(N) - 2\ln(\widehat{L})$, where k is the number of model parameters, N is the number of data points being fitted, and $\widehat{L} = p(D|\widehat{\theta}, M)$ is the likelihood corresponding to the parameters obtained by fitting; lower BIC scores correspond to better models. This quantity can be intuitively interpreted to counterbalance the quality of fit to the observed data - the $-2\ln(\widehat{L})$ term, and the model complexity - the $k\ln(N)$ term: compared to a model that is "just right", a too complex model is likely to produce a smaller

negative log likelihood value, but is penalised more for the larger number of parameters; a too simple model will have a higher value of the negative log likelihood. We discuss information criteria that we used in our analyses in more detail in the relevant sections (3.3 and 4.3).

Finally, paralelling the above discussion about parameter fitting, there are various approaches to model comparison for populations of subjects. When models have a hierarchical structure, model comparison is automatically performed at the population level. When fitting is performed for each subject individually, model comparison can proceed under the assumption that the same model is valid for all subjects, in which case evidence for each model is summed across the population, or under the assumption that different subjects can be characterised by different models, in which case model comparison involves determining model prevalence across the population (Stephan et al., 2009; Rigoux et al., 2014). We provide details of our implementation of model comparisons across our population of subjects in our analyses reports in following chapters.

Chapter 2

Experiment and task: concept, implementation, quantification challenges

In this chapter we provide a detailed presentation of the experiment from which the data analysed in subsequent chapters has been obtained. We begin by presenting our concept and goals, and the way they have informed our design of the experiment; we then present the task and the nature of the responses that we collected, briefly mentioning aspects of the data that constitute direct measures of the extent to which experimental manipulations were successful. Finally, we present some of the quantification challenges that we faced due to the nature and novelty of the task, and our approaches for dealing with them.

Our analyses suggested a number of task improvements that could be implemented in future work; we discuss them in the final chapter of this thesis (see section 6.2).

2.1 Concept, experiment goals

This experiment was conceived as an exploratory attempt to bring together two approaches in the study of attributions and the formation, maintenance and updating of beliefs about the self, which have complementary advantages.

One is the qualitative approach that has a long tradition of use in psychology and psychiatry (see review in chapter 1), which has the advantage of dealing directly with highly relevant concepts and questions, immediately meaningful on a human level; however these concepts are not easily amenable to the mathematical formalism that could provide a quantitative understanding of the underlying mechanisms. The other is the quantitative, often normative approach that the computational neuroscience and computational cognitive science fields have been more recently employing (e.g. (Behrens et al., 2007; FitzGerald et al., 2015; Daw et al., 2011), see also review in chapter 1); this has the advantage of elegantly designed, well controlled experiments, based on solid mathematical formulation of theories; however if often involves drastic simplifications of the concepts and phenomena of interest.

There are a number of criteria that we aimed for our task to satisfy: it needed to provide subjects with real, experienced outcomes, engaging enough to produce relevant attributions and relevant beliefs about the self; do so repeatedly, in order to allow for potential interactions between these two phenomena to manifest themselves; to control the subjects' experience and produce outcomes based on objective, measurable and controllable parameters; to involve learning, in order to allow for natural dynamics of belief updating and attributions. Finally, to enable investigation of whether these mechanisms are exclusive to the self, the task needed to be suitable for both actor and observer conditions.

2.2 Implementation

For these purposes, we implemented the task as a game of skill, providing subjects with an engaging and relevant context, and eliciting real attributions and beliefs about self. Subjects were asked to provide three kinds of responses: a causal attribution for the outcome they just experienced, an estimate of current skill and a bet on performance on the following trial. The task was presented to the subjects in two conditions, which we refer to as the "self" and "other"

condition, always presented in this order. In the "self" condition subjects played the game themselves and the attribution, skill report and bet questions referred to their own performance. In the "other" condition, subjects watched recorded trials, and answered the three types of questions about the recorded agent's performance. Subjects were told recorded performances belonged to another subject having previously completed the task, but were, in fact, presented with their own previous performance from the "self" condition.

2.2.1 The game

The task was implemented as a game of skill, inspired by the "Penguin Pursuit" game on the infamous Lumosity "brain training" platform `https://www.lumosity.com/en/`. While these games' claimed benefits on "brain power" are debatable, the games are undoubtedly engaging, challenging, motivating and entertaining. These are all qualities that we hoped would make the artificial lab context more relevant to our subjects, motivate them when performing the task and make them genuinely care about their performance, which is a prerequisite to obtain relevant, authentic attributions and beliefs about themselves.

In our visually simplified version of the game, subjects are presented, on every trial, with a maze they need to navigate through. They do so by using the arrow keys to guide a token red square from the starting position to the finish position marked by a trophy (see figure 2.1). If the red square reaches the goal in the limited allocated time, the win is indicated by the appearance of a smiley face; otherwise a frowning one appears to indicate loss.

The task is challenging because during each trial, repeatedly and unpredictably, the maze rotates and the correspondence between arrow keys and the direction of movement on the screen changes, according to the following rules. The maze is equipped with a "North" direction, marked on the screen by a compass needle; the arrow keys always move the red square toward the corresponding cardinal directions *of the maze* (Up towards "North", down towards "South" etc); initially, the maze's "North" correspond to the top of the

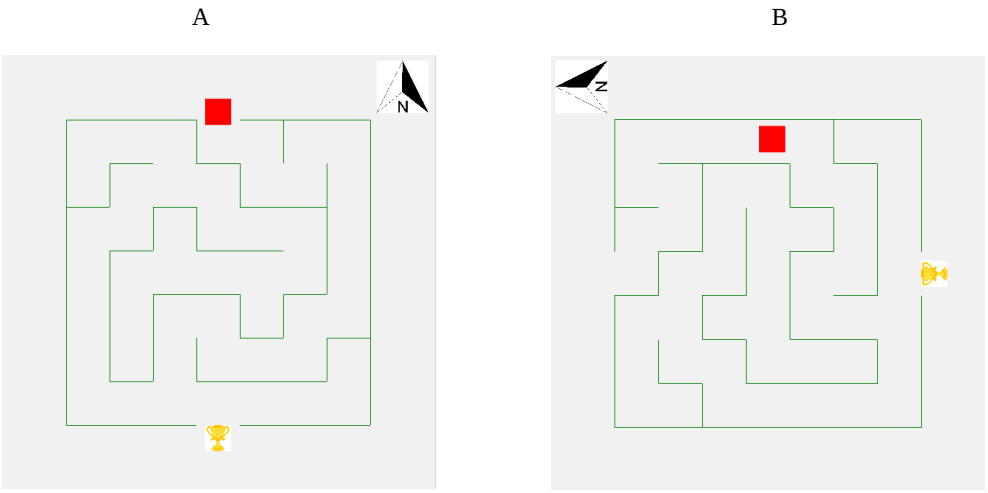


Figure 2.1: Task game illustration: two screenshots during one trial. Left: screenshot at the beginning of the trial (trials always start with mazes oriented upwards); pressing the up key moves the red square upwards on the screen, as indicated by the compass needle; the correct key to press in this case is down, moving the red square downwards on the screen. Right: later screenshot, the maze has rotated; pressing the up key now moves the red square towards the left on the screen, as indicated by the current orientation of the compass needle; the correct key to press in this case is also down, but it now results in moving the red square towards the right on the screen.

screen; however when the maze rotates, the compass needle rotates with it, such that "North" can point to any of the four directions on the screen, and accordingly pressing the up key no longer moves the red square up *on the screen*, but towards the maze's "North", wherever that is during each rotation. Subjects therefore have to learn to quickly adapt to the change in the correspondence between key presses and resulting movements on the screen.

Trial difficulty is controlled by a double staircase procedure, in order to prevent subjects from detecting the staircase pattern and having their responses or their performance during the trials affected by this knowledge. The staircase determining the parameters to be used is randomly chosen on every trial, the two stairs having equal probability, and is then updated according to the trial outcome. Both stairs have maze size, rotation frequency and available time as variables (see appendix F for a detailed description).The rather complicated stair structure and update mechanism encapsulates the in-

tuitive aspects of what constitutes difficulty in this game, as we were not in possession of a well established and calibrated one-dimensional measure of difficulty for this novel task (see sections 2.3.2 and 2.3.3 for an account of our work on extracting such one-dimensional measures of difficulty and objective skill from the data).

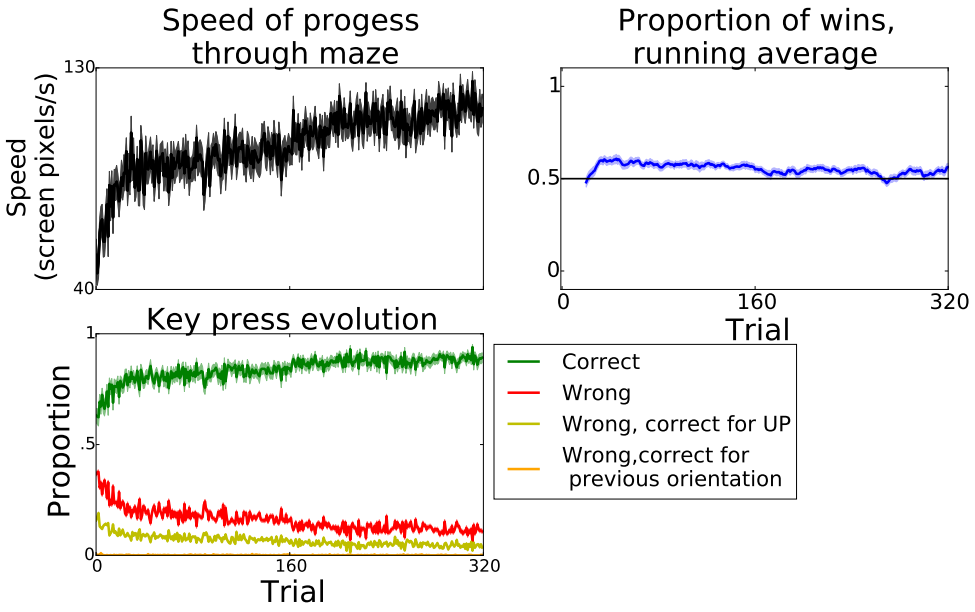


Figure 2.2: Evolution of performance and proportion of trials won, mean ± s.e.m. across subjects. Top left: speed of movement through the maze. Top right: proportion of trials won out of the last 20 trials. Bottom left: evolutions of proportions of key presses out of all key presses in every trial: types of key presses labeled as correct, wrong, wrong but correct for the normal UP orientation, wrong but correct for the previous maze orientation.

The practical goals of the staircase procedure were met, subjects displaying increasing performance, while also experiencing a relatively steady ratio of wins to losses (see figure 2.2). However, as discussed below (2.3.2 and 2.3.3) the procedure did not result in optimal adaptation to subjects' skill level, and its use inadvertently introduced confounds in the data, which we discuss in later sections dedicated to presenting data analysis and results (see 4.2.3.2).

2.2.2 Attributions, skill reports and bets

Questions about attributions and beliefs about skill were asked as follows: every two trials[1], immediately after feedback on the trial outcome, subjects were asked to provide a causal attribution for the outcome, to estimate their current skill in playing the game, and to bet on performance on the next trial, in this order. Subjects were informed of the time available to answer these questions (20 s), and that reward for the following trial would be withheld if they failed to provide an answer in this time limit.

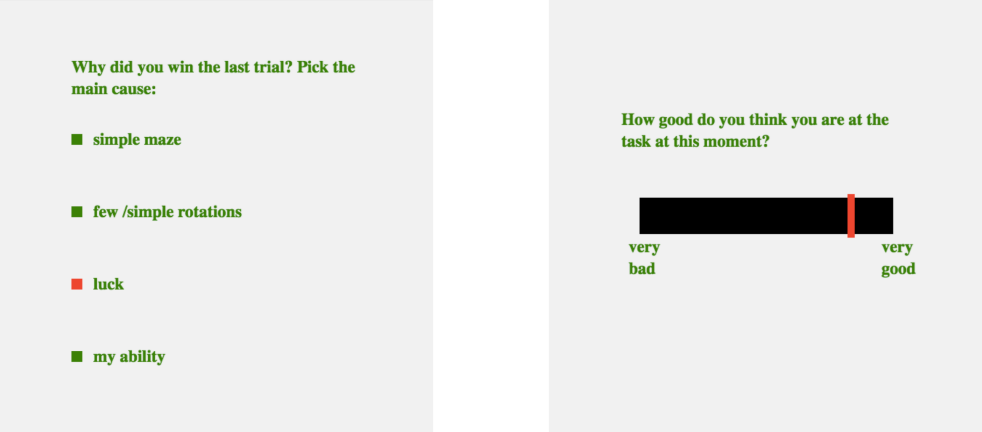


Figure 2.3: Left: Attribution question, self condition; Right: Skill question, self condition

Attributions were elicited with multiple-choice questions with the following response options(see figure 2.3): "simple maze", "few/ simple rotations", "luck", "my ability"[2]; the corresponding version for the ability option in the "other" condition was "their ability"[3]. These options reflected the internal vs external aspect of attribution ("self"/"other" vs other options), two different quantifiable parameters of the task (maze complexity vs rotations), and the option of blaming or crediting luck.

[1] We decided against asking these questions after every trial due to time constraints, and because the repetition might reduce subjects' attention and motivation in answering them. We note however that having measured these variables only every other trial introduced particular challenges for data analysis (see chapters 3 and 4 and discussion in 6).

[2] For losses, the options were "complex maze", "many/difficult rotations", "bad luck", "my lack of skill", respectively.

[3] And "their lack of skill" for losses.

This choice of response options was due to several factors: the internal-external distinction has been a central distinction in attribution theory and previous research (see review in chapter 1); the two specific task-related options allow quantification of relationships between attributions and objective task features; and causal explanations involving chance have been of interest in research on control and in studies of attributional patterns of psychiatric patients (Levenson, 1974; Kinderman and Bentall, 1996). However our choice of response options also introduced an undesirable availability bias toward external attributions (see discussion in sections 4.1 and 4.6.5).

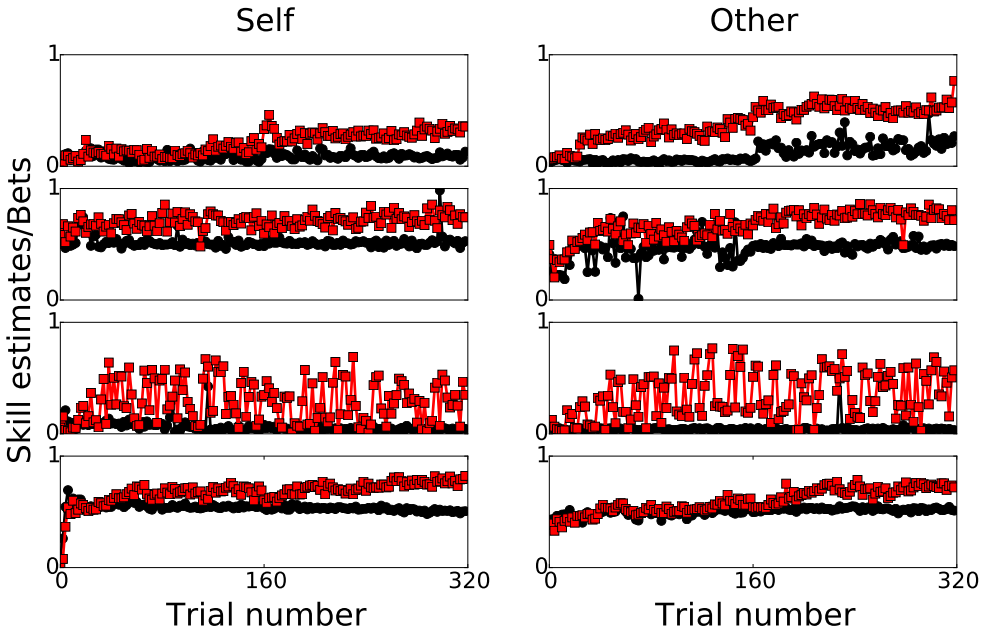


Figure 2.4: Bet vs skill responses, subjects settling on a constant betting response, in contrast with evolving skill responses. Red: skill responses; black: bet responses. Left: responses in the 'self' condition; right: responses in the 'other' condition. Subject ids, from top to bottom: 01060318, 01220218, 02210318, 04050318.

Subjects' beliefs about skill were probed with two questions. The first asked subjects to report how good they think they are (the "other" is) at the task by sliding a bar on a continuous scale between the "very bad" and "very good" extremes(see figure 2.3); responses were converted to values in $[0, 1]$.

The second question was aimed at eliciting an indirect assessment of skill, by asking subjects to bet on their (the "other"'s) performance on the

next trial. This was done through a Becker - DeGroot - Marschak (BDM) auction procedure (Becker et al., 1964), which guarantees that the optimal behaviour for the subject is to report their true estimate of the probability of winning (of the "other" winning) the next trial; this probability corresponds to the estimated skill level. The BDM procedure had been thoroughly explained to subjects prior to the start of the experimental session (see appendix A).

Note that while, in theory, subjects were expected to provide their true skill estimate as the BDM bet, in practice taking in the BDM procedure was challenging for most subjects. And while it is possible to verify that they understood it theoretically and were able to correctly answer questions about BDM auctions prior to the start of the experiment(see appendix A), in practice they might not access this knowledge while giving quick answers during the experiment, and might fall back on other patterns of betting. Indeed we found evidence of this in the data, with several subjects settling on a constant bet, unrelated to their concurrent skill responses(see figure 2.4). We have therefore discarded bets from subsequent analyses, and report only analyses of skill estimates in the rest of this work.

In contrast with the BDM procedure, there is no theoretical guarantee that the optimal behaviour is for subjects to answer the skill and attribution questions truthfully, nor any monetary incentive for them to do so (indeed the same is true for questionnaire responses).

There is, however, also no monetary incentive for them to respond falsely; their responses have no real influence on subsequent trials, and there is nothing in the instructions or in the task that might suggest a strategy of "gaining the system" by not replying truthfully. Note that this does not exclude other motivations to respond falsely, such as self-serving biases (see discussion in section 4.6 of chapter 4). The extent to which subjects answered these questions honestly in our case cannot be established, as there is no normative account of what the answers should have been, but our data indicates that subjects did not provide random answers even in the absence of enforcing

factors (see chapters 3, 4).

2.2.3 Conditions

As mentioned above, the experiment involved two conditions: first subjects played the game themselves and answered questions related to their performance - "self" condition; then subjects were instructed that they would watch trials recorded from another subject having previously participated in the experiment and they were asked to make attributions, evaluate skill and bet on this "other"'s performance - "other" condition. The claim about another subject was misleading: in fact, each subject was shown their own recorded trials from the "self" condition. In order to reduce the likelihood of participants recognising their own performance, mazes were left-right mirrored when they were played back.

Establishing the extent to which this manipulation succeeded is challenging, as questions to this effect risk alerting subjects to the deceit. Responses in the debriefing questionnaire (see appendix D) indicated that subjects did register similarities between their own and the "other"'s trials, however differences between responses in the two conditions were consistently present (see discussions in chapter 3 for skill and chapter 4 for attribution). Observed differences between responses in the two conditions can be due to the distinction between acting and watching, as well as to the distinction between self and non-self, and the present design cannot disentangle the contributions of these two factors. Improving this aspect of the experiment remains a goal for future work (see 4.6 and 6.2).

2.2.4 Payment

Subjects were payed a fixed amount for their time (£7.5 per hour) and they could gain additional monetary reward based on their performance: 5 trials were randomly picked out of trials in each session ("self" and "other"); subjects gained an additional £1 for each trial won out of the 5 selected; if selected trials included trials for which they had placed a bet, bet results were added

to subjects' earnings. Subjects were informed of the payment schedule prior to the start of the experiment (see appendixes A and B).

2.2.5 Questionnaires

In order to investigate relationships between behaviour in our task and well-established questionnaire-based measures of related psychological dimensions, we administered three questionnaires: the Attributional Style questionnaire (ASQ) (Peterson et al., 1982), the Levenson Locus of control scale (Levenson, 1974) and the Rosenberg Self esteem scale (Rosenberg, 1965) (see appendix E). Questionnaires were administered once, at the beginning of the experiment. We discuss the questionnaire responses and their relationship with the task responses in chapter 5.

2.2.6 Session timing

Pilot data indicated an average task duration of two hours per condition for a total of 320 trials, corresponding to 160 attribution and skill responses. To reduce the risk of subjects becoming fatigued or bored, each condition was split into two one hour sessions; these could be played (watched, in the "other" condition) either on the same day, with a few hours' pause, or on successive days, according to subjects' time availability. The resulting break between sessions was in some cases reflected in subjects' responses (see chapter 3).

2.2.7 Subjects

Subjects (n= 31, 9 males) were recruited through the UCL Institute for Cognitive Neuroscience (ICN) subject database[4] and were healthy, 18-35 year-olds (24 ± 3.9), fluent English speakers, with no history of neurological disorders. Subjects were required not to have used cannabis in the previous 31 days, not to have used any other recreational drugs in the week prior to their participation and not to have drank alcohol 24 hours before the study. The experiment was conducted under the ethics approval of UCL Departmental Research Ethics Committee, as Project ID Number fMRI/2013/005.

[4] `http://groupspaces.com/ICNSubjectDatabase/`

2.2.8 Experiment timeline

On their first lab visit, after signing the informed consent form, subjects completed the three questionnaires, on paper. They were then presented with the detailed instructions for the task, and answered test questions to ensure understanding of the game and BDM procedure (see appendix C). They then played 5 practice trials and the experimental trials began. The second session in the "self" condition started with experimental trials directly.

On their next visit a week later, subjects were provided with detailed instructions for the "other" condition (see appendix B) and then the first session in this condition started. The last session's experimental trials started directly.

After the end of the last session subjects were payed and provided with a feedback questions form, then with a separate question asking them to compare the two sessions (see appendix D). They were then fully debriefed.

2.3 Quantification challenges

Since our task is novel, there was no previously validated measure of either difficulty or skill.

Difficulty should integrate various objectively measurable features of the task into a one-dimensional score. It is important, since we expect it to be perceived by subjects and to affect their causal attributions and belief updating: winning a trial perceived as very simple might have less impact on a subject's belief about how good they are than winning a trial perceived as being very difficult.

Likewise, skill should combine different measurable aspects of performance into a one-dimensional objective score. Being able to objectively measure subjects' skill would be useful for establishing the degree to which their beliefs about their own skill were accurate and for investigating relationships between subjects' accuracy and their pattern of making attributions.

In this section we present our approach aimed at extracting such measures from the data, and analyses evaluating the quality of the measures we

obtained.

Difficulty and skill both contribute to determining outcomes, indeed to some extent they can be seen as opposite sides of a same coin; in addition, the staircase introduces dependencies between the evolution of skill and that of difficulty. It is therefore not obvious whether their contributions can be disentangled under these circumstances. We begin with presenting simulations which prove that recovering both skill and difficulty from data is possible, despite this recursive definitional complexity.

We then present the analyses we performed on the real data, and the resulting measures of difficulty and skill.

2.3.1 Simulations

Our simulations formalized a very simple case in which the outcome on trial t, $o(t)$, arose probabilistically from the interaction between an evolving scalar skill, $s(t)$, and an evolving scalar difficulty[5], $d(t)$, according to a Bernoulli distribution: $o(t) \sim \text{Bernoulli}(\sigma(s(t) - d(t))$[6].

Difficulty was generated as $d(t) = 1.7 f_d(t)$, a linear transformation of a 1-dimensional difficulty feature, f_d, whose evolution was determined by the simulated staircase[7]:

$$
\begin{aligned}
f_d(0) &= 0 \\
f_d(t+1) &= \begin{cases} f_d(t) + \delta(t) & \text{if } o(t) = 1 \\ f_d(t) - \delta(t) & \text{if } o(t) = 0, \end{cases} \\
&\text{where } \delta(t) \geq 0 \text{ is the staircase step on trial t.}
\end{aligned}
\tag{2.1}
$$

The staircase step $\delta(t)$ was sampled from a probability distribution, which

[5] We also simulated the case of two difficulty features, with similar results.

[6] Complexities in the recovery of difficulty and skill that arise in this simple framework will surely only be exacerbated in the real data. We conceived of these simulations as proof of concept for the validity of our approach.

[7] Note that the relationship between difficulty and the difficulty feature was fixed across subjects, as we assumed was the case in the real data, whereas the difficulty feature evolved for each subject as a function of their own outcomes, in accordance to the way the staircase worked in the real data.

varied among the different simulations, as detailed in the following sections.

In addition to time series of outcomes, $o(t)$, difficulty features, $f_d(t)$ and difficulty, $d(t)$, simulated data also included time series of performance features, $f_p(t)$ and skill, $s(t)$, which were generated in different ways in different simulations, as detailed below. We used the number of subjects and the number of trials of our real dataset.

In all simulations, the aim was to recover difficulty and skill – the variables which in our real data were latent – using the series of outcomes and the series of difficulty and performance features, which correspond to observable variables in the real data.

Difficulty and skill recovery was performed in the same way in all simulations. This was done according to the approach that we intended to use when extracting difficulty and skill from real data. Specifically, we would first obtain a measure of difficulty by exploiting the relationship between difficulty features and outcomes as revealed by time series of outcomes and task features from the whole population of subjects; this would thus capture a "general" measure of difficulty, valid for our population. We would then use this difficulty measure as a reference point with respect to which we would infer skill for each subject, using their own time series of outcomes and performance features[8].

Thus for all simulations difficulty was inferred first, as the linear transformation of the difficulty feature that best predicted outcomes in data pooled at the population level, according to the model $p(o(t) = 1) = \sigma(-d(t))$ on every trial. In order to exclude any effects of a subject's skill on the recovery of difficulty for their own trials, difficulty recovery for every subject was performed by fitting outcomes from all remaining subjects. Skill was then defined, for each subject, as the linear transformation of their time series of performance feature that best explained their outcomes while accounting for difficulty, according to the model $p(o(t) = 1) = \sigma(s(t) - d(t))$ on every trial.

[8]Note that other approaches, such as directly inverting the model including both skill and difficulty, are possible.

Skill was recovered independently for each subject, using the difficulty values recovered in the first step[9].

We performed three types of simulations, investigating the effect of several factors on the recovery of difficulty and skill, which we now present in turn.

2.3.1.1 Staircase step size and variability

The first factor that we examined concerned the step-size of the staircase. Because of the specific nature of our staircase procedure, we were worried that instead of closely tracking subjects' skill levels, the resulting difficulty changes were imprecise, leading to large difficulty steps from one trial to the next. We therefore first investigated the effect of the size of the staircase step on difficulty and skill recovery.

Difficulty changes generated by the staircase procedure in our task are also likely to vary significantly from trial to trial. This is due to several factors: the fact that we used two staircases, the hierarchical nature of the staircase changes, as well as the nature of the staircase levels, which involved available ranges, rather than unique feature values. We therefore also checked difficulty and skill recovery when large variability between step sizes was present.

For these simulations we assumed skill to be a linear transformation of a 1-dimensional performance feature, generated for each simulated subject as a 4-parameter sigmoidal function of time. The parameters of the sigmoid used to generate the performance feature, as well as those of the linear transformation from performance feature to skill, were chosen independently for each subject (see figure 2.5 for a plot of the resulting skill values).

The staircase step on every trial was drawn randomly from a narrow Gamma distribution; we compared the effects of a range of step sizes by using distributions with varying modes, ranging from 0.05 to 8.

[9]Note that because difficulty is obtained by predicting outcomes in the absence of skill, we expect difficulty to be recovered, at best, up to a scaling of the real values. The same is true for skill, as it is recovered with respect to the recovered difficulty. However the analyses involving objective difficulty and skill that we are interested in performing would not be impaired by such scaling.

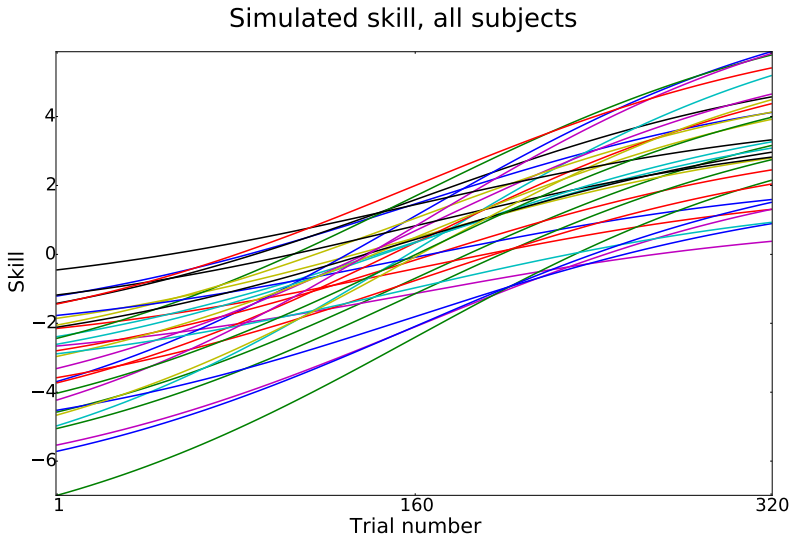


Figure 2.5: Simulated skill values. Each curve corresponds to one simulated subject.

We found that difficulty and skill recovery were successful – by which we mean difficulty and skill were recovered up to a scaling factor – for all but the extreme cases (see appendix G for details) of step size distributions (see figures 2.6 and 2.7 for an example, showing a summary of the simulated data and the quality of difficulty and skill recovery, respectively).

We also investigated the effect of using highly variable step sizes by generating staircase steps from an exponential distribution. We found that in this case too difficulty and skill recovery were successful (see figure G.4 in G). We used exponential distributions for all remaining simulation analyses.

We concluded from this set of simulations that difficulty and skill recovery work in this setup, for a large range of staircase updating parameters.

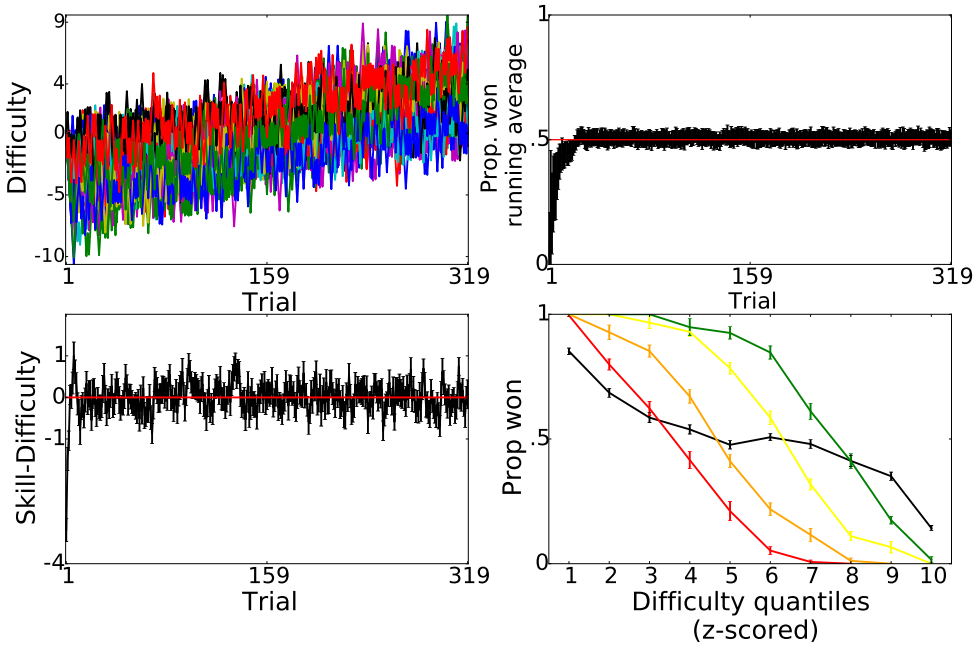


Figure 2.6: Example summary of simulated data for step size drawn from Gamma distribution (parameters used: $\theta = 0.007, k = 151$). Top left: difficulty evolution for all simulated subjects. Top right: running average of the proportion of trials won, mean $\pm$ s.e.m across subjects, filtering window: 20 trials. Bottom left: evolution of difference between skill and difficulty, mean $\pm$ s.e.m across subjects. Bottom right: relationship between difficulty and the proportion of wins, overall and as a function of skill level; mean $\pm$ s.e.m across subjects; black: overall, red to green: 1^{st} to 4^{th} skill quartiles; difficulty z-scored within subject.

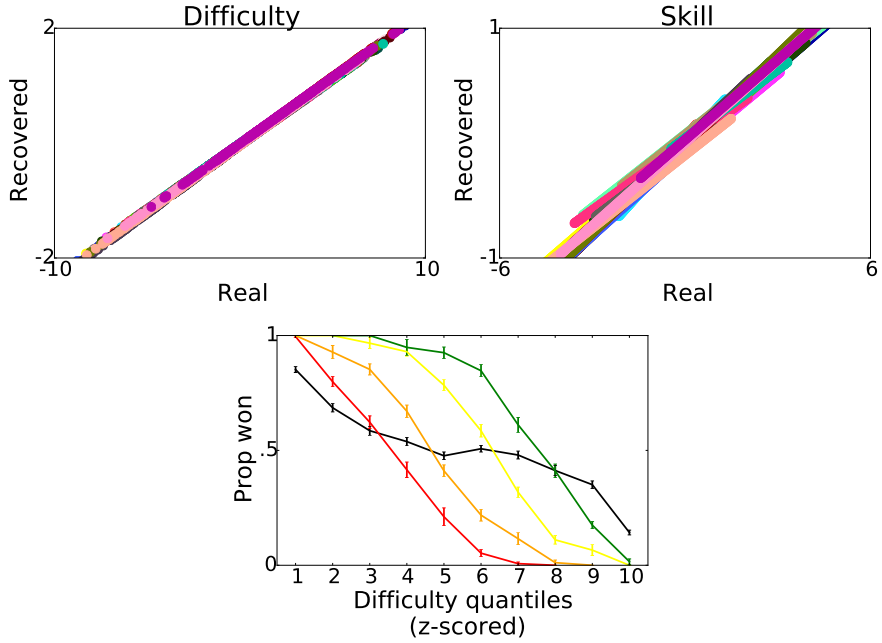


Figure 2.7: Difficulty and skill recovery for data in figure 2.6. Top difficulty and skill recovery, each color represents one simulated subject. Bottom: relationship between recovered difficulty and the proportion of wins, overall and as a function of recovered skill level; mean $\pm$ s.e.m across subjects; black: overall, red to green: 1^{st} to 4^{th} recovered skill quartiles; recovered difficulty z-scored within subject.

2.3.1.2 Non-monotonic skil

In the simulations described above, we used monotonically increasing performance feature profiles for all subjects. However this is not what we observed in the performance measures we collected from subjects. While performance and skill can be expected to have a tendency to improve throughout the task, local variations from this pattern are also likely, and there is no guarantee that subject's underlying skill monotonically increased throughout the task. We therefore performed a next set of simulations using performance features

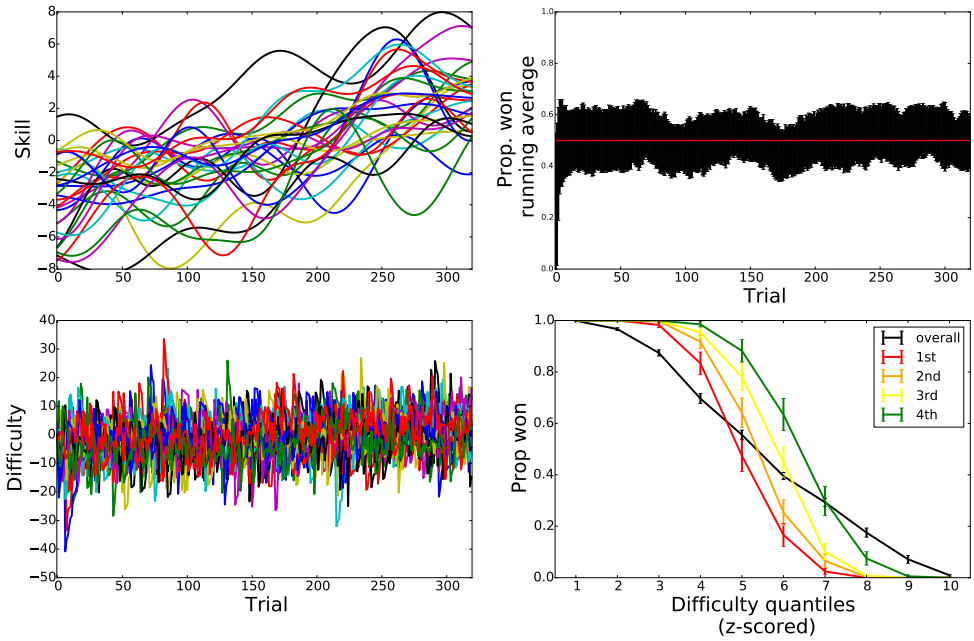


Figure 2.8: Example summary of simulated data, performance measures drawn from a Gaussian Process. Top left: skill evolutions for all simulated subjects. Top right: running average of the proportion of trials won, mean ± s.e.m across subjects, filtering window: 20 trials. Bottom left: evolution of difficulty for all simulated subjects. Bottom right: relationship between difficulty and the proportion of wins, overall and as a function of skill level; mean ± s.e.m across subjects; black: overall, red to green: 1^{st} to 4^{th} skill quartiles; difficulty z-scored within subject.

drawn from a Gaussian Process (GP) (Rasmussen and Williams, 2006) whose mean increased monotonically as a function of time. Specifically, we used the sigmoid performance feature profiles from previous simulations as the GP mean for each simulated subject and a Gaussian kernel. Skill measures were

linear transformations of the performance features, as in previous simulations. See figure 2.8 for a summary of the simulated data. We found difficulty and skill recovery to be successful in this case as well (see figure 2.9).

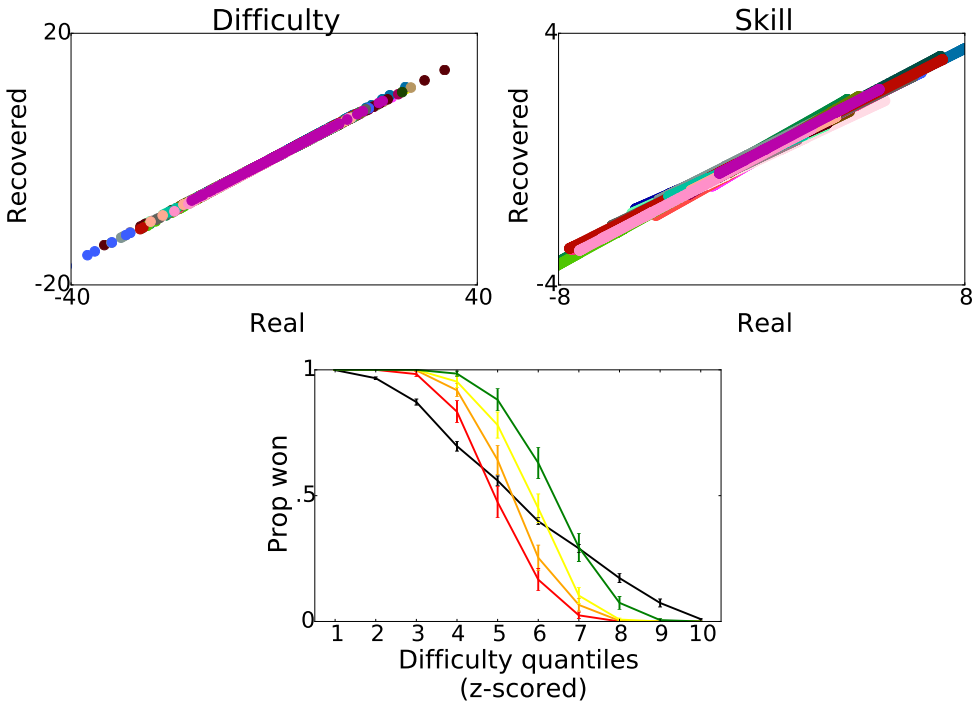


Figure 2.9: Difficulty and skill recovery for data in figure 2.8. Top: difficulty and skill recovery; each color represents one simulated subject. Bottom: relationship between recovered difficulty and the proportion of wins, overall and as a function of recovered skill level; mean $\pm$ s.e.m across subjects; black: overall, red to green: 1^{st} to 4^{th} recovered skill quartiles; recovered difficulty z-scored within subject.

2.3.1.3 Different generating and recovery models

In the simulations described previously, skill was generated as a linear transformation of the performance feature, and skill recovery proceeded under the same assumption. However this simplifying assumption might not be accurate, as it is likely that a more complex model underlies the relationships between task features, performance features, difficulty and skill in the real data. We therefore investigated the effect that using a different recovery model, instead of the real generating model, has on skill and difficulty recovery.

Rather than an integrated measure of performance, skill could also be conceived as a latent variable generating various observable, but noisy vari-

ables which constitute measurable performance features. The last type of simulation that we performed involved using this latter model to generate the data, while using the same procedure as in previous simulations for difficulty and skill recovery. Specifically, we used draws from a GP (with means generated as described before, and the same Gaussian kernel) as underlying measures of skill, and generated performance features as noisy linear transformations of these, for every simulated subject. Figure 2.10 shows a summary of the data obtained from one such simulation.

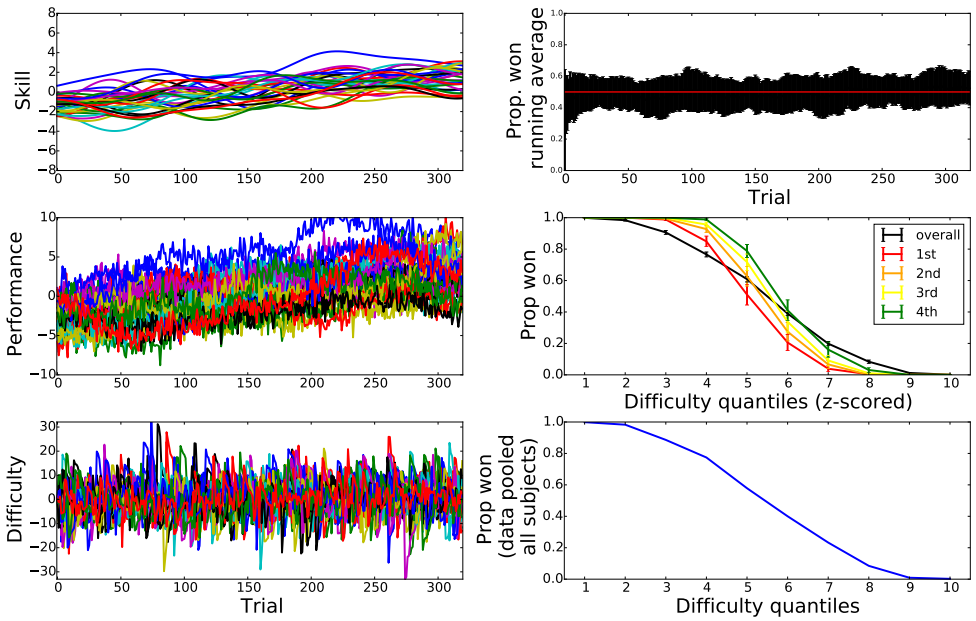


Figure 2.10: Summary of simulated data, skill measures drawn from a Gaussian Process, performance features as noisy linear transformations of skill. Left: evolutions of skill (top), performance(middle) and difficulty (bottom) for all simulated subjects. Right: Top: running average of the proportion of trials won, mean $\pm$ s.e.m across subjects, filtering window: 20 trials. Middle: relationship between difficulty and the proportion of wins, overall and as a function of skill level; mean $\pm$ s.e.m across subjects; black: overall, red to green: 1^{st} to 4^{th} skill quartiles; difficulty z-scored within subject. Bottom: relationship between difficulty and the proportion of wins, data pooled from all subjects.

For difficulty and skill recovery we used the same procedure as above; however as the recovered skill inherits the noisy nature of the performance feature, in this case we performed an additional filtering step, aimed at smoothing the recovered skill.

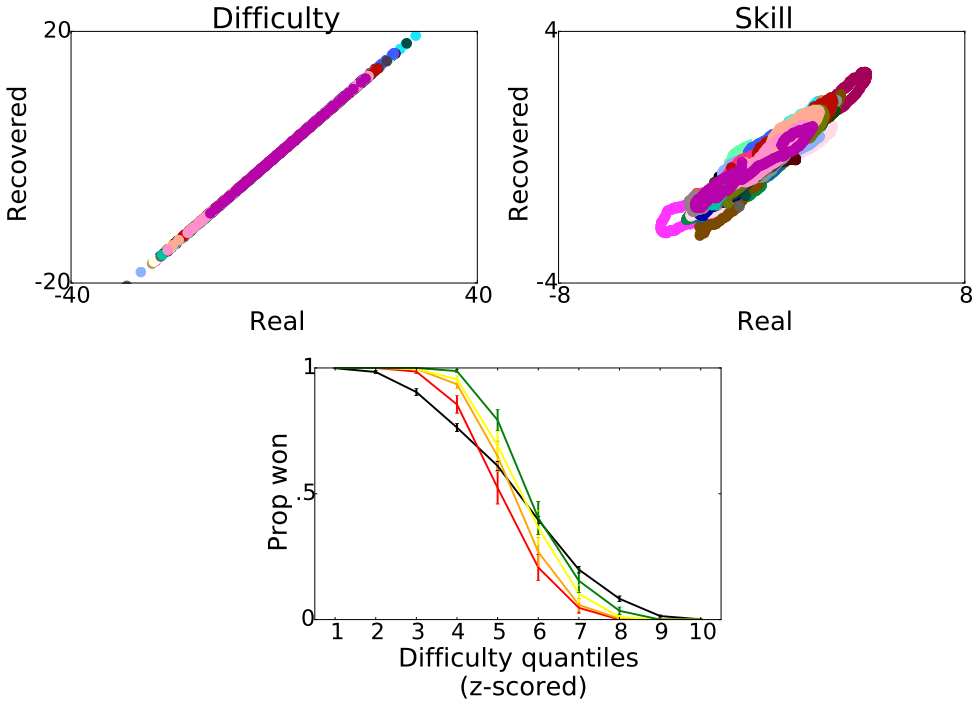


Figure 2.11: Difficulty and skill recovery for data in figure 2.10. Top: difficulty and skill recovery; each color represents one simulated subject. Bottom: relationship between recovered difficulty and the proportion of wins, overall and as a function of recovered skill level; mean $\pm$ s.e.m across subjects; black: overall, red to green: 1^{st} to 4^{th} recovered skill quartiles; recovered difficulty z-scored within subject.

We found that in this case too difficulty and skill could be recovered (see figure 2.11).

We concluded from these simulations that in this very simple setting, the interdependency between skill and difficulty introduced by the use of a staircase does not necessarily prevent recovery of both difficulty and skill from the data, in the absence of external calibration. This does not guarantee successful recovery of difficulty and skill in our data, which involves complexities that were not present in our simulations – difficulty and performance features were not multi-dimensional and, more importantly, dependencies might exist between difficulty and performance features– however it constitutes a proof of concept for the validity of our approach. We therefore proceeded to define

difficulty and skill post hoc in our real data, which we describe next[10].

2.3.2 Empirical difficulty

As mentioned above, our strategy for extracting objective measures of both difficulty and skill from our data was to first establish a measure of difficulty, and then define skill with respect to it. Intuitively, this corresponds to the following cartoon scenario: a difficult trial is one that a randomly picked subject in the population is likely to lose, while an easy trial is one that subjects are likely to win; subjects' skill is measured with respect to this difficulty - the more likely subjects are to win difficult trials, the more skilled they are considered to be. In this section we present the construction of the difficulty measure.

The staircase controls several objective dimensions along which trials vary, and which could contribute to their objective difficulty(see figure 2.12). Different factors might be weighted differently in an integrated difficulty score, and we had no a priori knowledge about these individual contributions. Our approach, described below, consisted in inferring the identities and weights of the different factors from data, i.e. from subjects' performance.

In order to do so, we made a number of assumptions: specifically, we assumed difficulty to be a linear combination of objectively measurable factors, with unknown weights; we further assumed the factors' weights to be stable across time; finally, we assumed that the resulting difficulty measure predicts trial outcome, allowing weights to be inferred from outcome prediction.

Specifically, given a set of objectively measurable task aspects that might act as factors in determining difficulty, $f_1, f_2, ..f_k$, we assumed that there is a stable set of weights, $w_1, w_2, .., w_k$, representing the contribution that they

[10]Note that in all simulation analyses we assumed the relationships between difficulty features and difficulty, as well as those between performance feature and skill, to be stable across time. This is a key ingredient of the success of difficulty and skill recovery, and an assumptions that we made in analyses of real data. It is an important limitation of our approach; future work producing larger data sets that could be used for calibration would remove the necessity for this assumption.

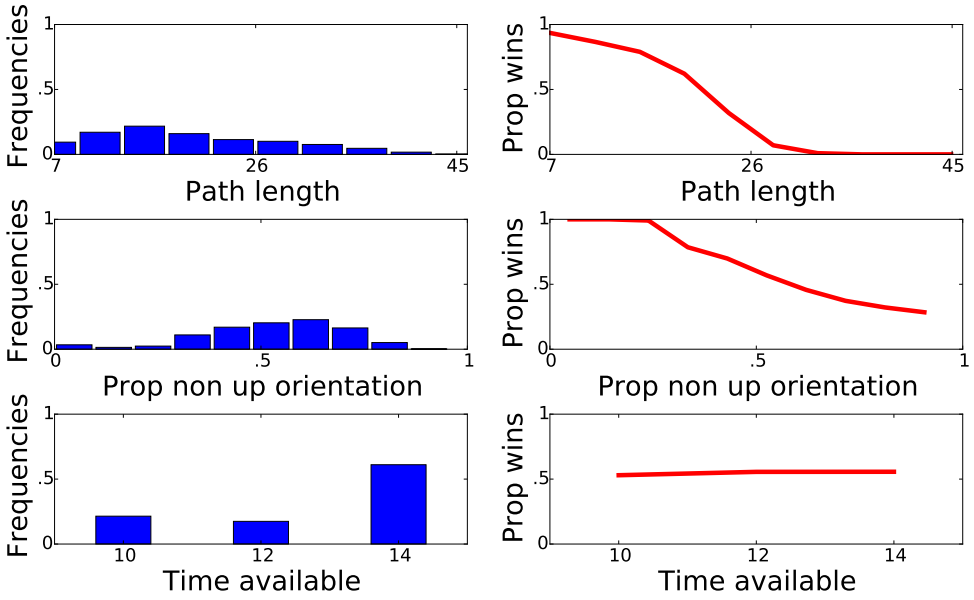


Figure 2.12: Objective task features controlled by the staircase procedure and their relationship with outcomes; data pooled from all subjects. Left: frequency histograms of the three features. Right: proportion of trials won as a function of the three features. Top: length of correct path through the maze. Middle: proportion of maze orientations within a trial different from the normal UP orientation. Bottom: available time.

each have towards determining the outcome, such that on any trial t

$$
\begin{aligned}
p(o(t)=1;\mathbf{w}) &= \sigma(\mathbf{w}_d^T\mathbf{f}_d(t)), \text{where} \\
\mathbf{w}_d^T &= (w_0, w_1, \ldots w_k) \\
\mathbf{f}_d^T(t) &= (1, f_1(t), \ldots f_k(t)) \\
f_i(t) &= \text{the measured value of factor } f_i \text{ at trial t} \\
o(t) &= \text{outcome at trial t, and} \\
\sigma(x) &= \frac{1}{1+\exp(-x)}, \text{the sigmoid function.}
\end{aligned}
$$

Given $\mathbf{w}_d$, difficulty for trial t can be computed as

$$d(t) = -\mathbf{w}_d^T\mathbf{f}_d(t). \tag{2.2}$$

Under these assumptions, obtaining a measure of difficulty is equivalent to identifying $\mathbf{w}_d$, which can be inferred from outcome prediction. In order

to obtain, for each subject, an objective, external difficulty measure[11], not influenced by their own skill, we inferred $\mathbf{w}_d$ separately for each subject, by predicting outcomes for all remaining subjects[12].

In order to choose among the measurable aspects of the task that contribute to difficulty, we fitted all outcomes from all subjects with logistic regression models with different combinations of regressors and their interactions. We compared 9 models based on combinations of the following regressors: length of correct path to maze exit, proportion of non UP orientations, time available, necessary minimum speed (itself a combination of path length and time available). See figure 2.12 for the distribution of the basic factors (length of correct path to maze exit, frequency of rotations, time available) and their relationship with outcomes.

Based on the cross-validation score, the model which best explained outcomes across subjects using only objective information about the trial included length of correct path to maze exit, proportion of non UP orientations, path length$\times$orientation interaction, time available and necessary minimum speed as features. We therefore used the same regressors for building a working measure of difficulty for each subject.

The distribution of difficulties we obtained satisfies the two properties that we used as sanity checks:

- the probability of winning decreases with increasing difficulty across subjects
- subjects' difficulty indifference points(the difficulty for which they are

[11] We also computed an alternative, 'individual' difficulty measure for each subject, using exclusively information from their own trials. In this case weights for each subject were inferred by predicting their own outcomes only. As this measure uses information from all trials, it is still unrealistic as a 'subjective' measure : subjects do not have access to information from later trials when perceiving difficulty at any given time during the experiment. The 'individual' and external measures of difficulty we obtained were different, but highly correlated, with an average correlation coefficient $\rho = 0.98$. Unless otherwise stated, in the rest of this work 'difficulty' will be used to refer to the external, objective measure of difficulty.

[12] We compared two ways of integrating data from remaining subjects, pooling subjects together and modelling variability between subjects with a hierarchical model, with very similar results; see appendix H for details. In the rest of this work we will use the difficulty measure obtained by pooling.

equally likely to win or lose the trial) correlate with overall performance in terms of their total number of wins and losses.

Indeed as can be seen from figure 2.13, a sigmoid shaped relationship exists between difficulty and the proportion of wins vs losses across subjects. This relationship is also preserved at the individual subject level; in addition, if a sigmoid is fitted to each subject to predict outcome based on difficulty, the inflexion point of the fitted sigmoid, which constitutes the difficulty value for which the subject is equally likely to win or lose the trial, correlates with the subject's overall proportion of wins.

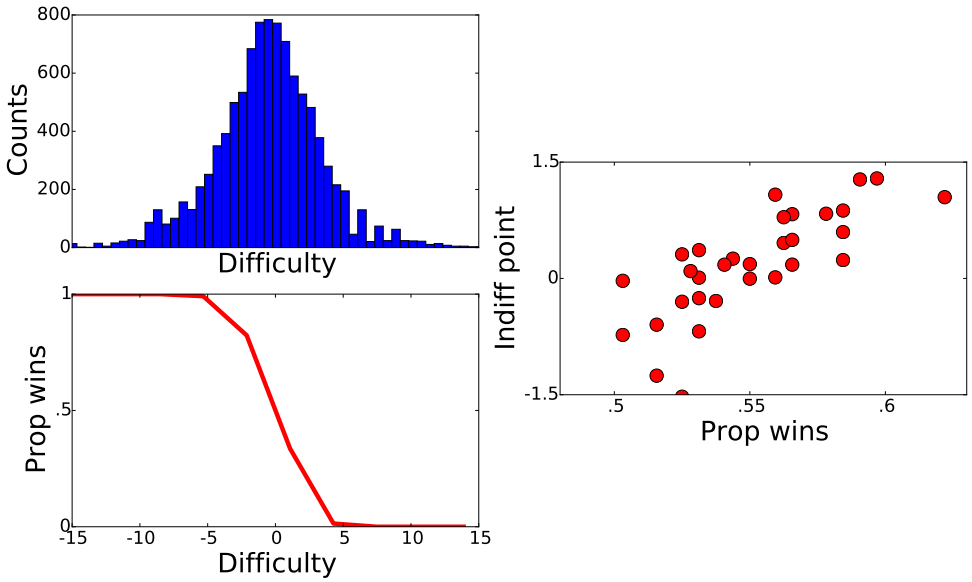


Figure 2.13: Difficulty measure: summary and sanity check. Left: data pooled from all subjects; top: distribution of difficulty values; bottom: relationship between difficulty and the proportion of wins. Right: relationship between difficulty indifference point - difficulty value for which subject is equally likely to win or lose the trial - and the proportion of trials won out of all trials; each dot represents a subject; $r^2 = 0.6$, p-value $= 3 * 10^{-7}$.

Due to the adaptive stimulus presentation, we expected difficulty to increase over time as subjects got better at the task. We therefore also tested wether this is the case for the difficulty measure we obtained, by comparing the average difficulty of the first and last quarter of trials across subjects. We found that the average difficulty indeed increases for most subjects(2 sample

t test $t = -4.78$, p-value $= 10^{-5}$). In addition, difficulty variability also increases (s.d comparison between first and last quarter of trials: $t = -6.42$, p-value $= 2 * 10^{-8}$); this is an undesired result of the way we designed the step changes in our staircase procedure: the staircase steps did not control difficulty directly, but only changed maze dimensions, average rotation frequency and time available. Thus, rather than directly shifting the distribution of difficulty up or down, the effect of a staircase step was in some cases to extend (or contract) the range of difficulties available. Consider for instance a step up on the staircase consisting in a larger available maze size; this increases the likelihood of mazes with long paths to exit, but very simple mazes with a short, direct path to exit are still possible, even though more unlikely; the result is therefore a broadening of the range of possible difficulties.

This aspect of the staircase design is not ideal, and could be improved in future work: an optimal functioning of the staircase would track the subject's skill, such that trials are neither too difficult, not too easy. Analysis of the accuracy of difficulty as a predictor of outcome suggests that the staircase might indeed have failed to track subjects' skill levels: outcome prediction based only on difficulty values, according to the simple model $p(o(t) = 1) = \sigma(-d(t))$, is highly accurate, ranging from 0.69 to 0.91, with an average of 0.84 and s.d. of 0.04 across subjects(see figure 2.14). Tracking the subject's skill level would explore difficulty ranges where difficulty alone is insufficient as a predictor(see figure G.2 in appendix G). The fact that difficulty alone is such a good predictor of outcomes implies that little room is left for skill; this is indeed what we found in our analyses aimed at defining skill, to which we turn next.

2.3.3 Empirical skill

We consider objective skill to be the evolving factor that intermediates between the objective difficulty of a trial and success. We therefore sought a validated way of integrating our various objective measurementsof subjects' performance into this latent construct. We attempted to extract an objective

measure of skill from data, following an approach similar to the one we employed for difficulty. We note, however, that due to factors discussed below, we were not able to obtain a satisfactory measure of skill, and therefore did not use it in further analyses as initially planned.

Our approach to defining objective skill relied on the following assumptions: we assumed skill to be a linear combination of objectively measurable performance features, with unknown weights; these weights we assumed to be stable across time; finally we assumed a linear combination of objective difficulty (as previously defined) and skill predicts trial outcome. In this setup, identifying objective skill is equivalent to finding the linear combination of measured performance features that can best predict trial outcomes over and above difficulty. Formally, this corresponds to the following model for outcomes:

$$
\begin{aligned}
p(o(t)=1;\mathbf{w}_p) &= \sigma(\mathbf{w}_p^T\mathbf{f}_p(t)+\mathbf{w}_d(t)),\ \text{where}\\
o(t) &= \text{ outcome of trial t}\\
\mathbf{f}_p(t) &= \text{vector of performance regressors at trial t}\\
d(t) &= \text{ difficulty at trial t}\\
\mathbf{w}_p &= \text{performance weights, parameters}\\
\mathbf{w}_d &= \text{difficulty weight, fixed}
\end{aligned}
$$

For a given value of $\mathbf{w}_d$, fitting the above model to trials from one subject produces performance weights $\mathbf{w}_p$; these can then be used to obtain the trial by trial skill measure, computed as the performance contribution to the outcome prediction:

$$s(t) = \mathbf{w}_p\mathbf{f}_p(t)$$

We used three performance features, computed on a trial-by-trial basis, namely the proportion of pauses, the proportion of correct key presses, and the proportion of wrong key presses that would have been correct in the normal

UP orientation[13]. Finally, as the recovered skill inherits the noisy nature of the performance features, we performed an additional filtering step, aimed at smoothing the recovered skill.

We compared the model's prediction accuracy for a range of negative values for $\mathbf{w}_d$, as well as for $\mathbf{w}_d = 0$, which is equivalent to using only performance features to predict outcomes (see figure 2.14); we also compared these accuracies with that obtained when using difficulty as the only predictor of outcomes. Note that the accuracy of the models including performance features is a training accuracy (and therefore likely an overestimation), as these models were fitted on the same individual subject data on which accuracy was computed; this is not the case for the difficulty-only model (see section 2.3.2).

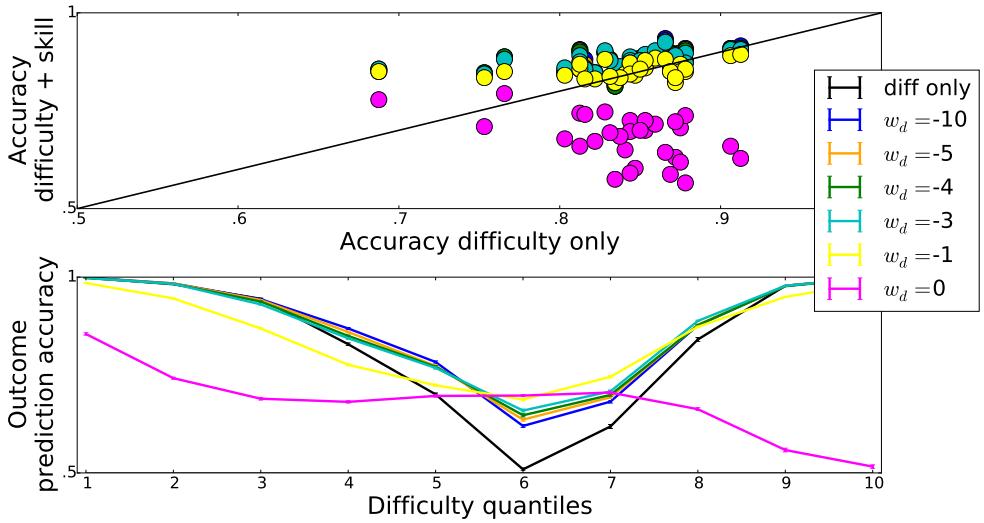


Figure 2.14: Accuracy for outcome prediction: difficulty only vs difficulty and skill models. Colours correspond to different $\mathbf{w}_d$ values in the skill and difficulty models; note that $\mathbf{w}_d = 0$ (purple) is equivalent to a model with skill only; black is used for the model with difficulty only. Top: overall accuracy; each dot represents one subject. Bottom: accuracy per difficulty level; mean $\pm$ s.e.m across subjects; difficulty was z-scored for each subject and discretised in 10-quantiles.

As illustrated in figure 2.14, this comparison showed that difficulty alone

[13]Note that this trial-wise summary represents a coarse view of subject's performance, which, given the complexity of task, presents reach within-trial dynamics. See appendix I for more details.

is overall more predictive of outcome than performance features alone, and that adding performance features to the difficulty-only model only marginally improves overall accuracy. Computing outcome prediction accuracy as a function of difficulty level provides a more detailed account of the models' performance, showing that there is only a narrow range of difficulty values for which difficulty alone fails to predict outcome, and where including performance features significantly improves prediction accuracy.

This observation indicates that the staircase procedure did not sufficiently adapt difficulty to subjects' skill level, exploring, instead, more extreme difficulty ranges. Since it is mainly trials with intermediate levels of difficulty that are informative about subjects' skill levels, it is difficult to establish to what extent our assumption of a constant relationship between performance features and skill is valid in our data.

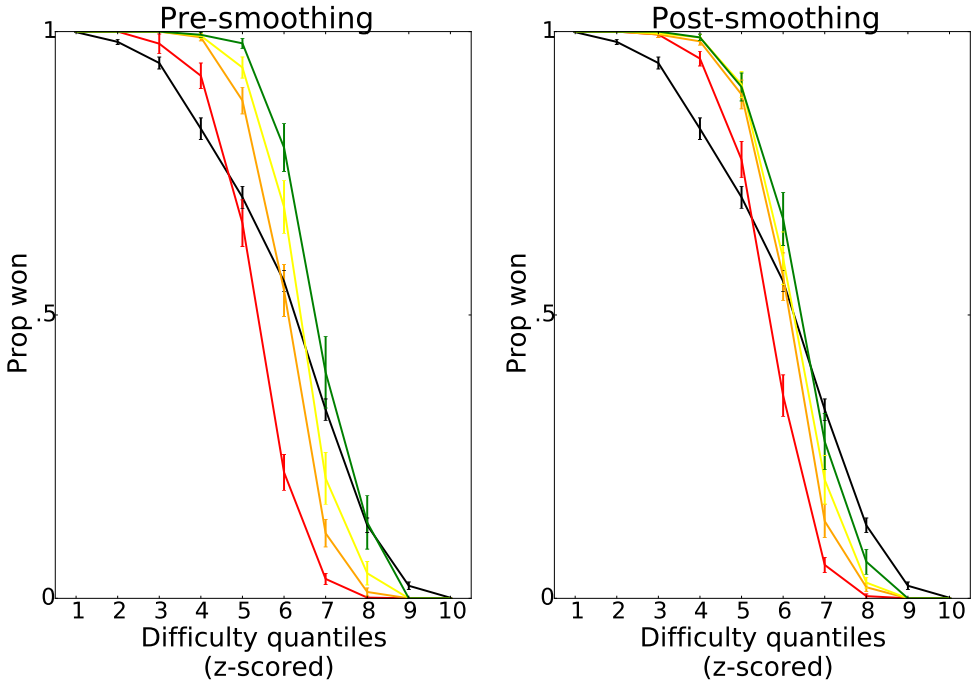


Figure 2.15: Relationship between difficulty and the proportion of wins, overall and as a function of skill level; mean ± s.e.m across subjects; black: overall, red to green: 1^{st} to 4^{th} skill quartiles; recovered difficulty z-scored within subject. Left: extracted skill measure before smoothing. Right: extracted skill measure after smoothing, filtering window size = 5 trials.

We used the effect of skill on the relationship between difficulty and out-

comes as a sanity check of the resulting skill measure: for a valid measure of skill, increasing skill would shift the sigmoid-shaped curve relating difficulty with the probability of winning towards the right, such that the higher the skill, the higher the difficulty level for which $p(win) = 0.5$ (see simulation summary figures in 2.3.1, e.g. 2.10 bottom right) .

This pattern was not convincingly displayed by the skill measured we obtained. Specifically, while the integrated performance measure resulting from the above skill recovery procedure displayed the expected pattern, the smoothing step dampened the effect (see figure 2.15). This was the case even when narrow filtering windows were used, presumably due to high levels of trial-to-trial variability in the raw (non-smoothed) integrated performance measure. We did not encounter this phenomenon in our simulations using noisy performance features (see 2.3.1.3), even when using large amounts of noise. This difference might be due to the fact that unlike simulated performance features, which were generated independently of difficulty, the real performance features that we used in the above analyses might reflect influences of difficulty as well as skill, and thus include additional trial-to-trial variability resulting from large staircase steps.

Due to these issues, we did not to use the skill measure resulting from analyses presented in this section in our subsequent analyses. However relationships between objective reality and subjects' subjective assessment of their skill are of particular interest for the study of self-beliefs and their relationships with attributions, and improving the staircase procedure to allow for a more accurate tracking of subjects' skill remains an important goal for future work (see 6.2).

Chapter 3

Skill estimates

Beliefs about skill and causal attributions are the two fundamental variables of interest in this study. In this chapter, we present our analyses of subjects' responses to the questions asking them to estimate their own/the "other"'s skill.

The chapter is structured as follows. The first section contains a summary presentation of the relevant data. The second and third sections present our analyses of skill estimates - model agnostic analyses in the second section and model-dependent analyses in the third; in both types of analyses effects of attributions on skill estimates were of particular interest, but we also investigated relationships between the evolution of skill estimates and several other variables of interest. The fourth section contains analyses of subjects' reaction times for providing answers to the skill questions. The chapter concludes with a summary and discussion.

3.1 Data summary

The data of interest in this chapter are skill estimates - the estimates provided by subjects for their own/the "other"'s skill, and skill updates - differences between successive skill estimates.

Figure 3.1 shows the evolution over trials of skill estimates in the "self" and "other" conditions, for all subjects; figure 3.2 shows the distribution of skill estimates provided in each condition, for all subjects. These figures il-

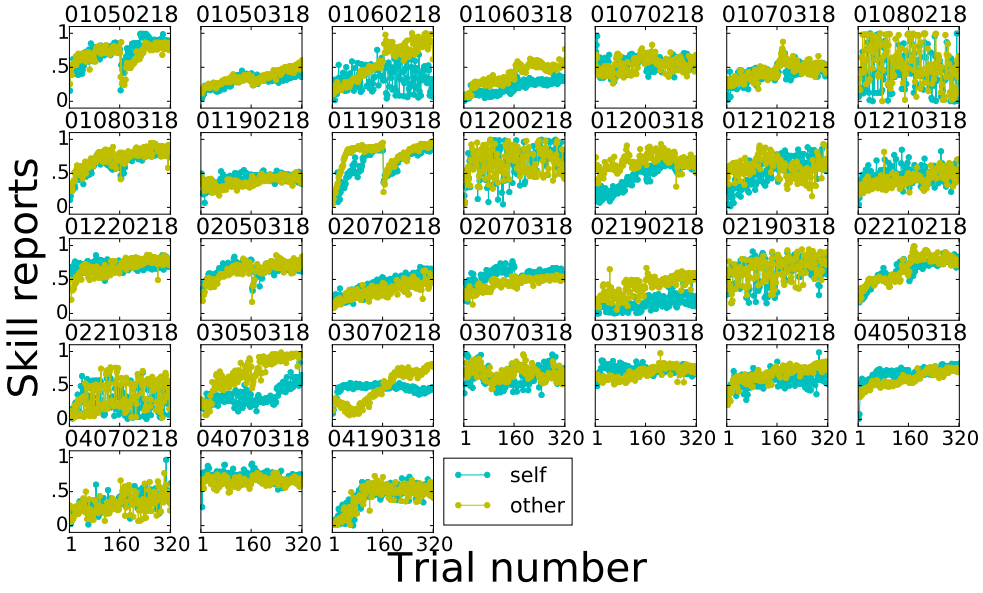


Figure 3.1: Time evolution of skill estimates, all subjects, both conditions.

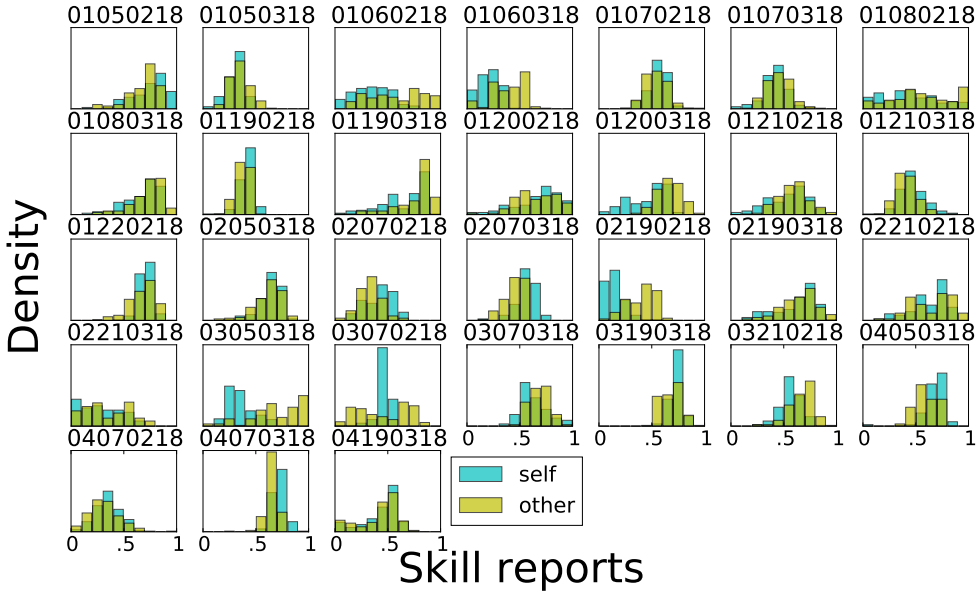


Figure 3.2: Distributions of skill estimates for self and other, all subjects. Overlapping density histograms.

lustrate the fact that there are common patterns in subjects' responses, but also significant variability between subjects.

In most cases, skill estimates for "self" and "other" appear to follow very similar trajectories, but there are subjects for whom the "self" and "other" skill estimates follow visibly different patterns (e.g. in figure 3.1, 01060218: row 1, column 3, 03070218: row 4, column 3, 03050318: row 4, column 2, and 01200318: row 2, column 5).

In general, skill estimates increase gradually, but there are also cases of large up and down variations between successive estimates (e.g. 01080218: row 1, column 7 and 02190318: row 3, column 6); there are also cases (e.g. 01050218: row 1, column 1 and 01190318: row 2, column 3) where there is a drop in skill estimates corresponding to the break between sessions(see chapter 2).

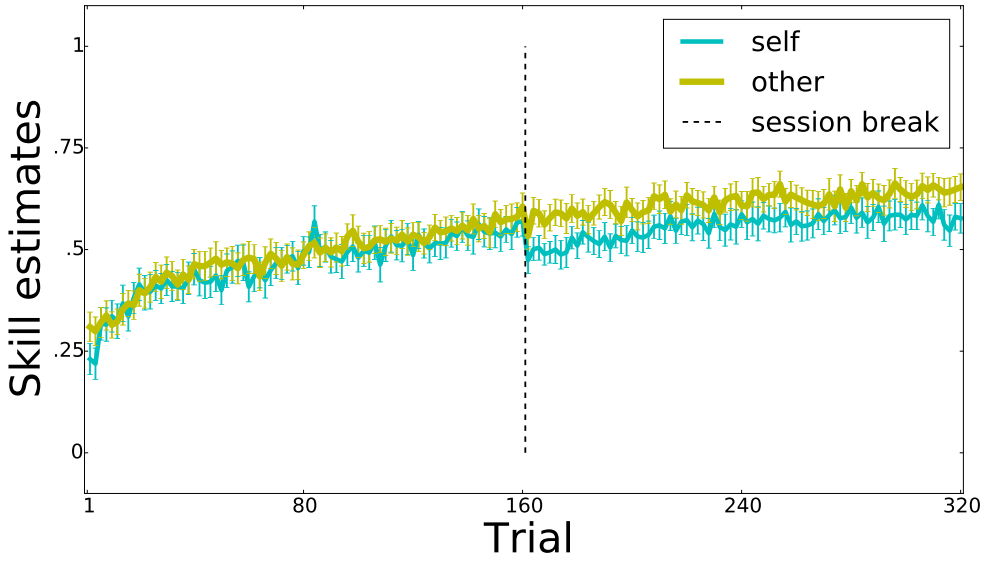


Figure 3.3: Skill estimates for self and other averaged over all subjects ± s.e.m

Trial by trial averaging over subjects shows that both the session break effect and the difference between "self" and "other" are present across subjects (see figure 3.3). Skill estimates for "other" were on average higher than for "self", particularly in the second session, despite the fact that subjects' answers to our debriefing questions (D) showed that subjects had some suspi-

cion that they were watching, at least partially, their own performance. This pattern of apparent bias in favour of the "other" is consistent with what we observed in subjects' attribution responses (see section 4.2.2 in chapter 4) and with observations from previous research in various contexts (Crockett et al., 2014; Rand et al., 2014) (see Rand and Nowak, 2013, for a review). See below and sections 4.2.2, 4.6 for analysis and discussion of this effect.

Figure 3.4 shows the distributions of skill updates for "self" and "other" for all subjects. With few exceptions (e.g. 01080218: row 1, column 7, 01200218: row 2, column 4, 02190318: row 3, column 6, 02210318: row 4, column 1) - updates were narrowly distributed around 0. However, the distribution across subjects of the average skill update is shifted to the right of 0 both for "self" and for "other" ("self": 1 sample $t(30) = 7.28, p < 0.01$, "other": 1 sample $t(30) = 6.79, p < 0.01$). This is consistent with the fact that skill estimates displayed learning curve-like trajectories at the individual subject level (see figure 3.1) and with the fact that, averaged over subjects, skill estimates increased over time (figure 3.3).

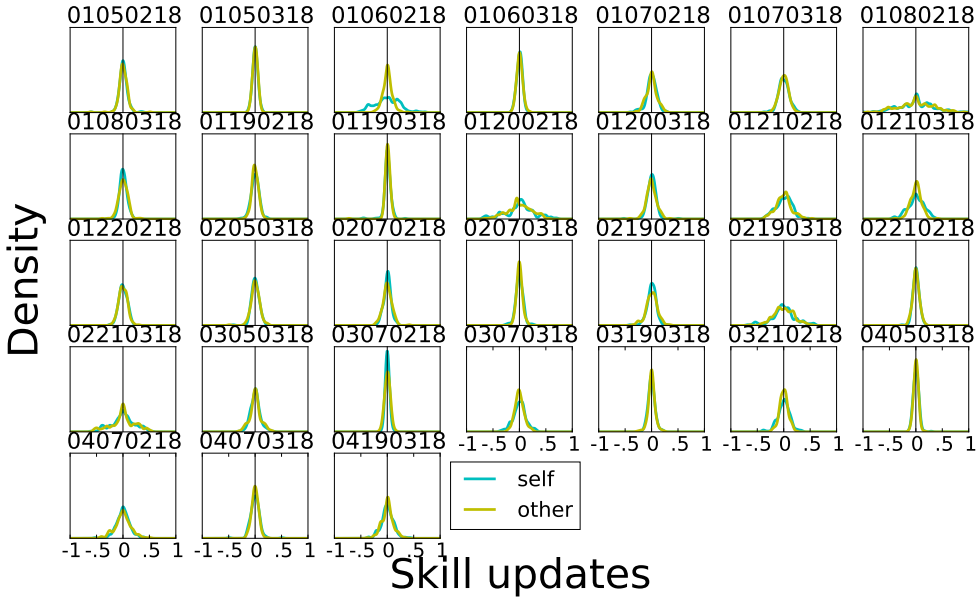


Figure 3.4: Distribution of skill updates, all subjects. Densities approximated with a Gaussian kernel density approximation, bandwidth = 0.03.

In order to investigate the mechanisms driving the evolution of skill es-

timates, and in particular their relationship with attributions, we performed both model-agnostic and model-based analyses. In the next section we present model-agnostic analyses.

3.2 Model agnostic analyses

There are four main factors that we expected would influence the evolution of skill estimates: trial outcome, trial difficulty, the subject's performance and the subject's causal attribution for the outcome. In this section we present model-agnostic analyses of skill updates, which we performed in order to determine whether these factors have detectable effects on a trial-by-trial basis.

We tested for the effects of interest as follows: for discrete variables (outcome and attribution), we computed the average skill update corresponding to each level of the factor of interest for every subject and tested for differences between the resulting distributions across subjects; for continuous variables (difficulty, performance measures) we computed the correlation between skill updates and the variable of interest for every subject, then tested whether the resulting distribution across subjects is shifted with respect to 0.

All continuous variables were z-scored within subject; skill updates were computed as variations in the z-scored values. Attribution options were relabelled as "internal" vs "external", as this was the distinction of interest. We applied the Benjamini-Hochberg procedure for multiple comparisons (Benjamini and Hochberg, 1995) to all but the post hoc tests, and all results reported as significant survived the correction procedure.

Note that due to our decision to only ask subjects for a skill estimate every two trials (see 2), we did not have access to the finest granularity of skill updates, but only to the aggregate effect of pairs of trials.

3.2.1 Outcome

Figure 3.5 illustrates the effect of the immediately preceding outcome, and the effect of the pair of outcomes from the previous skill estimate, on skill updates.

As expected, we found that skill updates immediately following a win are significantly higher than updates following a loss, for both "self" and "other" ("self": paired $t(15) = 6.43, p < 0.01$, Hedge's corrected d = 2.24, "other": paired $t(15) = 6.08, p < 0.01$, d = 2.12).

Pairs of outcomes experienced between skill estimates fall in one of 4 categories: (loss, loss), (win, loss), (loss, win), (win, win). Comparisons between the distributions of skill updates corresponding to the 4 categories also confirmed our expectations. These distributions were ordered as expected, indicating that the latest outcome carried more weight than the outcome one trial back, but also that the first outcome in the pair did matter ("self": WW vs LW paired $t(15) = 5.19, p < 0.01$, d = 0.87, LW vs WL paired $t(15) = 4.14, p < 0.01$, d = 1.34, WL vs LL paired $t(15) = 3.58, p < 0.01$, d = 0.49; "other": WW vs LW paired $t(15) = 2.6, p = 0.01$, d = 0.47, LW vs WL paired $t(15) = 3.87, p < 0.01$, d = 1.27, WL vs LL paired $t(15) = 3.13, p < 0.01$, d = 0.58).

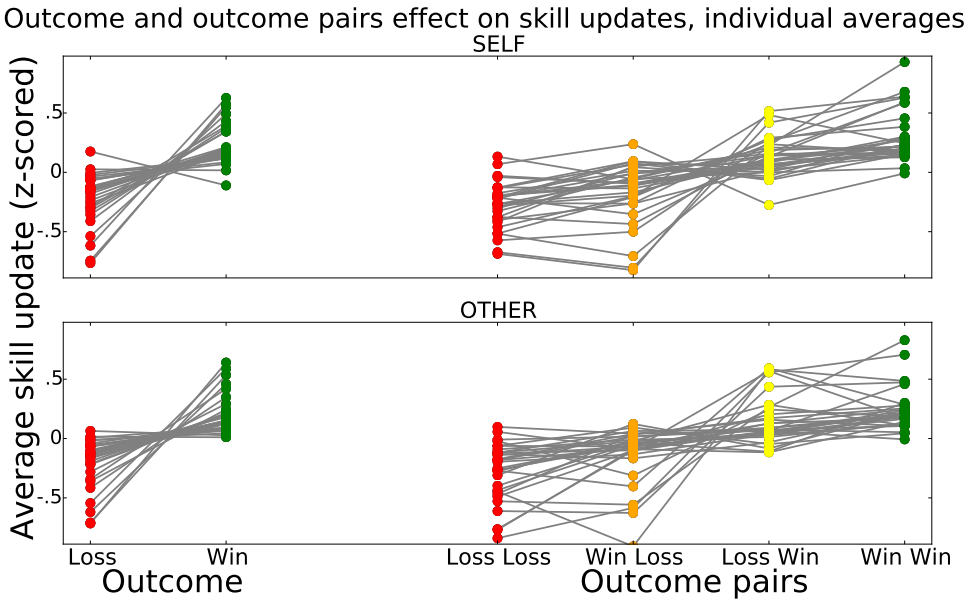


Figure 3.5: Distributions of skill updates, subject-level average z-scored data. Left: updates labelled according to immediately preceding outcome. Right: updates labelled according to the pair of outcomes experienced from previous skill report. Each dot represents a subject.

3.2.2 Attribution

We hypothesised that trials attributed internally (to self in the "self" condition, to other in the "other" condition) would contribute more to skill updating than trials attributed externally(to task aspects or luck) and that attributions would interact with the outcome, such that wins attributed internally would improve skill estimates more than wins attributed to the task, and losses attributed internally would reduce skill estimates more than losses attributed externally.

We performed 2-way repeated measures ANOVA (Howell, 2012) (see appendix O for details), with outcome (win vs loss) and attribution (internal vs external) as fixed within subject factors, and subject as a random factor (see figure 3.6)[1]. For the "self" condition we found no significant main effect of attribution, but a significant interaction with outcome ($F_{A(1,28)} = 1.62$, $p = 0.21$, $F_{AxO(1,28)} = 9.05$, $p < 0.01$). For the "other" condition, we found both a significant main effect of attribution, and significant interaction with outcome ($F_{A(1,31)} = 10.46$, $p < 0.01$, $F_{AxO(1,31)} = 16.51$, $p < 0.01$). In both cases the ANOVA also identified the significant main effect of outcome ("self": $F_{O(1,28)} = 35.35$, $p < 0.01$, "other": $F_{O(1,31)} = 38.47$, $p < 0.01$).

We performed post hoc repeated measures t-tests to test for a secondary effect of attribution conditioned on outcome. In the "self" condition we found no significant effect of attribution for wins, but a significant effect for losses ("self" internal vs external attributions: wins paired $t(14) = 1.25, p = 0.12$, d = 0.23; losses paired $t(14) = -2.45, p = 0.01$, d = -0.45). In the "other" condition we found significant effects for both wins and losses (wins: paired $t(15) = 2.19, p = 0.02$, d = 0.41, losses paired $t(15) = -4.46, p < 0.01$, d = -0.89). Direct comparisons of the absolute values of the effects for wins vs losses revealed that the effect for losses was significantly stronger in both conditions (wins vs losses "self": paired $t(14) = -1.89, p = 0.03$, "other": paired $t(15) = -2.54, p = 0.01$).

[1] For these analyses we excluded from the "self" condition 2 subjects who provided no internal attribution for wins.

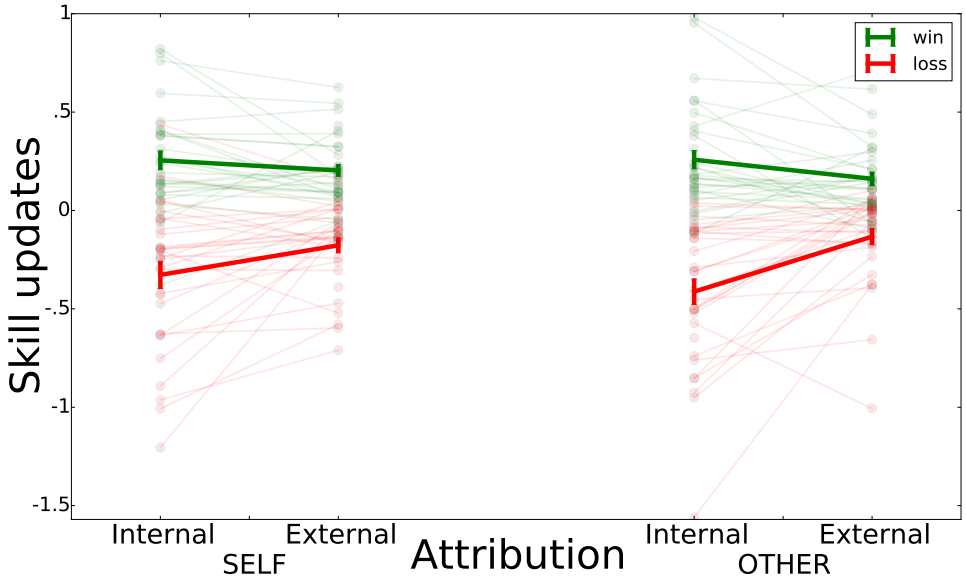


Figure 3.6: Attribution and outcome effects on skill updates, within-subject average data. Each dot represents a subject. Bold lines represent mean ± s.e.m. across subjects.

Finally, we performed direct comparisons between the strength of effects of attribution in the two conditions. We found no significant difference either for wins or for losses ("self" vs "other" wins: paired $t(14) = -0.89, p = 0.2$, losses: paired $t(14) = -1.5, p = 0.07$). However in all of the above analyses effect sizes were higher for "other" than for "self", which suggests a difference might be present; a larger data set is needed to convincingly determine whether the effect is real.

To conclude, these analyses provided evidence for the hypothesised interaction between attributions and outcomes, with secondary effects of attribution conditioned on outcome in the expected directions. Unexpectedly, we found the effect of attribution to be stronger for losses than for wins and possibly stronger for "other" than for "self". These observations provide testable hypotheses for future work, which we discuss in sections 3.6 and 6.3.

3.2.3 Difficulty

We hypothesised that difficulty interacts with outcome in influencing skill updates, such that winning a difficult trial is perceived to reveal higher skill

than winning an easy one, and therefore would lead to a larger increase in skill estimates; conversely losing a difficult trial would be less indicative of low skill than losing an easy trial, and would therefore lead to a smaller decrease in skill estimates.

Since our task is novel, we did not have a previously validated measure of difficulty, and instead used a difficulty measure constructed from our data, as discussed in 2.3.2. One important analytical caveat for assessing the relationship between difficulty and skill estimates is that difficulty was constructed so as to best predict outcomes, and outcomes themselves were associated with changes in skill estimates. We therefore checked for an effect of difficulty over and above outcome by investigating the relationship between difficulty and skill updates separately for wins and losses.

We computed, for each subject, the Spearman correlation between difficulty and skill updates, and tested whether the resulting distributions across subjects are shifted with respect to 0. We used the t-test statistic, but estimated p-values from permutation tests. We found a significant effect of difficulty on skill updates for wins("self": 1 sample $t(30) = 2.79, p < 0.01$, "other": 1 sample $t(30) = 2.92, p < 0.01$), but not for losses("self": 1 sample $t(30) = -0.49, p = 0.68$, "other": 1 sample $t(30) = 0.5, p = 0.31$). Direct comparison between the effects for the two outcomes identified the difference as significant for "self", but not for "other" ("self": paired $t(15) = 2.08, p = 0.03$, "other": paired $t(15) = 1.49, p = 0.07$).

Direct comparison between the strength of the effect of difficulty between "self" and "other" found no significant difference for either wins (paired $t(15) = -0.86, p = 0.21$) or losses (paired $t(15) = -0.19, p = 0.43$).

Thus an effect of difficulty on skill updates could be identified in the data, over and above the effect of outcome. The analyses described above also suggest there might be differences between processing difficulty for wins and losses, as well as between the "self" and "other" conditions, however evidence in this respect is inconclusive. There are a number of factors that

future work might improve on, allowing for a clearer picture of the effects of difficulty to emerge (see sections 3.6 and 6.3).

3.2.4 Performance

The last factor whose effect on skill updates we investigated in a model-agnostic manner concerns the details of performance, separate from the outcome. We expected better performance to be associated with higher skill updates than lower performance, for both wins and losses.

As was the case with difficulty, we did not have a previously validated integrated 1-dimensional measure of performance [2] Instead, we tested for effects of three directly measurable performance variables on skill updates, namely the proportion of correct presses out of all keys pressed during the trial (pc), the proportion of time spent pausing during the trial (pp) , and the proportion of wrong key presses which would be correct for the normal UP orientation (pwcu). This last performance factor we chose because it is one that could be particularly salient to subjects due to the nature of the task.

For each of these performance measures, we computed, for each subject, the Spearman correlation between the performance measure and skill updates, separately for wins and losses. We then tested whether the resulting distributions across subjects are shifted with respect to 0. We used the t-test statistic, but estimated p-values from permutation tests.

We found significant effects of all three performance measures on skill updates, in the expected direction[3], for losses, but not wins. This was true in both conditions (see table 3.1). Direct comparison between the effects for the two outcomes identified the differences between the effects for wins and losses as significant, with one exception (see table 3.1).

We further directly compared the strength of the effects for losses in the "self" and "other" conditions, and found no significant difference for pc (self

[2] As discussed in 2.3.3, our attempt to extract an objective measure of skill from data produced results that were not satisfactory.

[3] Thus positive correlation between skill updates and pc, and negative correlations between skill updates and pp and pwcu.

	Self		Other	
	t	p	t	p
pc +	0.58	0.29	0.64	0.26
pp +	1	0.16	-0.73	0.77
pwcu +	0.97	0.16	0.48	0.32
pc -	3.99	<0.01	3.8	<0.01
pp -	-2.86	<0.01	-2.1	0.02
pwcu -	-3.12	<0.01	-2.5	0.01
pc + vs -	-3.53	<0.01	-3.14	<0.01
pp +vs -	2.8	<0.01	1.17	0.12
pwcu + vs-	3.47	<0.01	2.71	<0.01

Table 3.1: Effect of performance measures (see text and glossary for abbreviations) on skill updates. Top: testing whether the distribution across subjects of the Spearman correlation coefficients between performance and skill updates for wins (+) and losses (-) is shifted with respect to 0. Bottom: direct comparison- testing whether the distribution of differences in correlation coefficients between wins and losses is shifted with respect to 0. One sample t-test statistics; p-values estimated from permutation tests.

vs other paired $t(15) = 1.65, p = 0.06$), and significant differences for pp (paired $t(15) = 1.81, p = 0.04$) and pwcu (paired $t(15) = 1.91, p = 0.03$).

The difference we found between wins and losses was unexpected, but could be explained by the existence of an adaptive mechanism involving higher discrimination in learning from losses than from wins. It is consistent with our observations on the effect of attribution, and might reflect a focus on learning from negative outcomes, in order to avoid them in the future, while allocating less discriminative attention to positive outcomes, which can more generally be used as a positive signal. The fact that the effect of performance on skill updates after losses seems to be stronger for "self" than for "other" is consistent with this view.

3.2.5 Model agnostic analyses: summary

Model agnostic analyses presented in this section identified significant effects of outcome, attribution, difficulty and performance on skill updates in the directions we had hypothesised. They also provided evidence for unexpected differences between these factors' relationships with skill updates post wins vs post losses, which suggest further questions for future work (see discussion

in 6.3).

Form an evolutionary perspective it is arguably more costly for an agent to make wrong inferences about their ability in negative situations than it is in positive ones. If a negative outcome is experienced, it is highly relevant for an agent to determine whether the failure is due to its own actions, which can be improved in order to avoid costly outcomes, or to the environment, in which case other cost-minimising strategies could be pursued. In contrast, inaccurately assigning responsibility for a positive outcome is unlikely to incur a high cost(although it could lead to lesser gains). This view is consistent with differences we observed between the effects of performance and attribution on skill updates post wins vs losses.

While this perspective could be relevant for explaining and further investigating differences between processing of positive and negative outcomes in the "self" condition (see discussion in section 3.6), it cannot explain why such differences would also exist in the "other" condition; it is, however, consistent with a scenario in which mechanisms used for updating beliefs about self and those used for others share common elements.

Effects of attribution were generally larger for "other" than for "self", suggesting attributions might be processed differently in the two conditions, however future work is needed to determine whether this is indeed the case. If so, there are a number of factors that could explain such differences: belief updating processes in the "self" condition might be more complex that processes in the "other" condition, and people might be more honest when reporting attributions and or skill estimates for "other" than they are when reporting on their own performance, which is presumably more emotionally salient and engaging. The difference between acting and watching might also be responsible for observed differences between the "self" and "other" conditions: in the "other" condition, subjects need only watch and evaluate , while they need to also act, and do so under time pressure, in the "self" condition. Task manipulations could be designed to tease these candidate explanations

apart (see discussions in section 3.6 and 6.3).

Finally, it is worth emphasising two important differences between the nature of the attributions involved in our analyses and that generally investigated in previous research(see review in 1): previous research often manipulated attributions by providing subjects with different cover stories in different conditions; in our case attributions were actually reported by subjects, on a trial-wise basis. Thus the above analyses provided evidence for an effect of attribution on beliefs at a higher temporal resolution than that explored by previous research; and they provided evidence for an effect of directly expressed individual attributions, rather than ones postulated based on condition-wise manipulations across subjects.

These analyses provided a population-level view of the factors' effects on skill updates. However they did not account for the time structure of the responses, nor for possible individual differences in mechanisms driving skill updates. In order to obtain a finer grained understanding of the data at the individual subject level, in particular taking into account the effect of time, we next turned to model-dependent analyses, which we present in the next section.

3.3 Model-dependent analyses

The aim of our model-dependent analyses was to investigate in more detail the relationships between the evolution of skill estimates and the factors of interest, particularly attribution, while accounting for the time-series nature of our data and for potential individual differences between subjects.

The modelling approach to data analysis relies on encoding hypotheses about the mechanisms underlying the observed data into precise mathematical formulations; these are used to generate data sets according to different candidate models and compare them to real data. The extent to which different models account for the observed data is used to evaluate the corresponding hypotheses that the models encode (see section 1.4 for a detailed account of

the theory). Ideally, different models vary significantly in how well they are able to account for the data, and model comparison is highly informative.

In particular, we expected model comparison to be useful in answering two types of questions: the first was to determine the contribution of individual factors of interest by comparing the quality of models which included them with that of models which did not; the second involved comparing models with different relationships between the factors on interest in order to study interactions between them. We compared a large number of models, falling under three main categories: purely descriptive models of the time evolution of skill estimates (see 3.3.1), Rescorla-Wagner models (Rescorla, 1972) (see 3.3.2), and "observing the observer" models (Daunizeau et al., 2010b,a) (see 3.3.3). The models included components which had different sources: some where directly related to our hypotheses (e.g. attribution-dependent learning rates), some were inspired by the results of our model-agnostic analyses (e.g. outcome modulation of attribution-dependent learning rates), others were made necessary by practical aspects of our data collecting process (e.g. effect of session break).

Despite between-subject variability in skill estimates, we expected the underlying mechanisms to be at least partially shared. In the modelling framework this would correspond to data from substantial numbers of different subjects being generated from the same model (or similar models), albeit with different individual parameters. In this case, model comparison would show patterns of preference between models that would be consistent for different subjects, thus revealing mechanisms valid at the population level (see section 1.4 for a detailed discussion of ways to account for variability between subjects in model comparison). Indeed, while subject-level answers to model-comparison questions are informative, their general relevance is naturally related to the extent to which they are valid across subjects and can thus generalise.

This, however, is not what we found in our case: both in terms of in-

formative differences between models, and in terms of the extent to which model comparison identified common mechanisms between subjects, results of model-dependent analyses differed from out expectations.

Thus, rather than observing significant differences between different models, we found that some subjects were poorly fitted by all models, while many of them were quite well fitted by almost all models (see figure 3.7 which, setting aside, for the time being, the issue of penalising for complexity, shows the pattern of r^2 scores obtained by all models for all subjects in the "self" condition). None of these patterns is as informative as we had originally hoped.

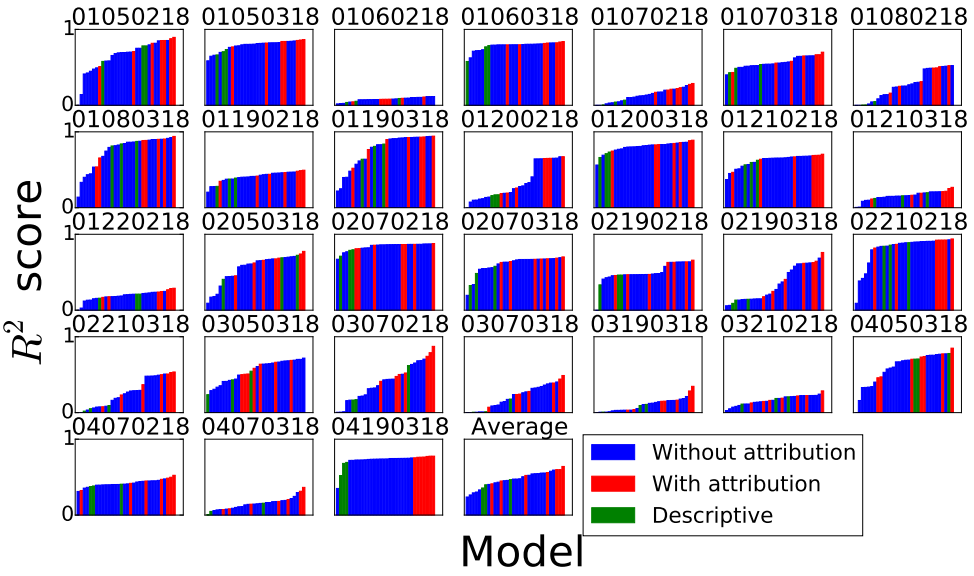


Figure 3.7: R^2 scores for descriptive and Rescorla-Wagner models for skill estimates, 'self' condition. Each bar represents a model, descriptive models are coloured in green, RW models without attribution in blue, RW models with attribution in red. Note that many of the subjects were either poorly fitted by all models, or quite well fitted by all models. Note that no complexity correction has been applied.

In addition, contrary to our expectations, variability between subjects was present not only in their patterns of skill estimates, but also in the models which best accounted for this data (see section 3.3.4).

Because variability between models at the subject level was lower than expected, we attempted to understand and clarify this more complex picture

of the data by gradually expanding our model family. There were two main mechanisms driving the expansion of the model family: one involved building increasingly complex models in order to test wether more complex relationships involving the factors of interest were able to better explain our data; the other involved testing a number of models that could alternatively explain the data without relying on the factors of interest.

Because variability between subjects in terms of preferred models was higher than expected, we did not attempt to draw any general conclusion about the belief updating mechanisms at the level of the population. The small number of subjects available also precluded any attempt to investigate the existence and nature of clusters in the population. Instead, we adopted a different approach, focusing on the individual subject level in order to determine whether there are any subjects for whom the factors of interest are significant (see 3.3.4). This strategy is relatively unconventional, however complex tasks can be expected to engage multiple mechanisms that are exploited in different ways by different individuals, and the strategy we have adopted aimed to find statistically-convincing evidence for such mechanisms.

This section is structured as follows: we begin by presenting the models we compared, with a focus on the processes driving the expansion of the model family (subsections 3.3.1 - 3.3.3); we then present our approach for performing model comparison, and its results (subsection 3.3.4). The section ends with a summary (subsection 3.3.5).

3.3.1 Purely descriptive models

As noted previously, for most subjects the evolution of skill estimates seemed to follow a learning-curve like profile, with a steep increase at the beginning followed by levelling out in a plateau. In order to measure to what extent incorporating hypotheses about the mechanisms driving skill updates enables models to fit the data better than purely descriptive models lacking any mechanistic insights, we fitted a small number of purely descriptive models to our data.

We used variations of two common learning curve models: a logarithm and a sigmoid shaped ones; models based on the two curves were labeled L and S respectively. The variations we introduced were meant to account for aspects of the data that our experimental procedure might have introduced.

Our model agnostic analyses revealed that the division of each condition into two separate sessions produced an unexpected jump in some of our subjects' skill estimates, between the last response from the first session and the first one from the second session. We refer to it as the session break effect (see figure 3.3). One model variation was therefore whether models did or did not have a parameter to account for this effect.

The existence of the break effect in some subjects also points to the fact that belief updating mechanisms might proceed differently between the two sessions. The second variation was therefore whether models were allowed to use different parameters for the two sessions, or were limited to the same parameters across sessions.

These variations led to 3 models for each underlying basic curve: no effect of session break (coded L1 or S1), effect of session break but same parameters across sessions (coded L2 or S2), effect of session break and different parameters for the two sessions (coded L3 or S3). The list of fully specified models can be found in appendix J.1.

Figures 3.8 and 3.9 show the quality of model S3 fit to data for two example subjects - the best fit and worst fit subjects.

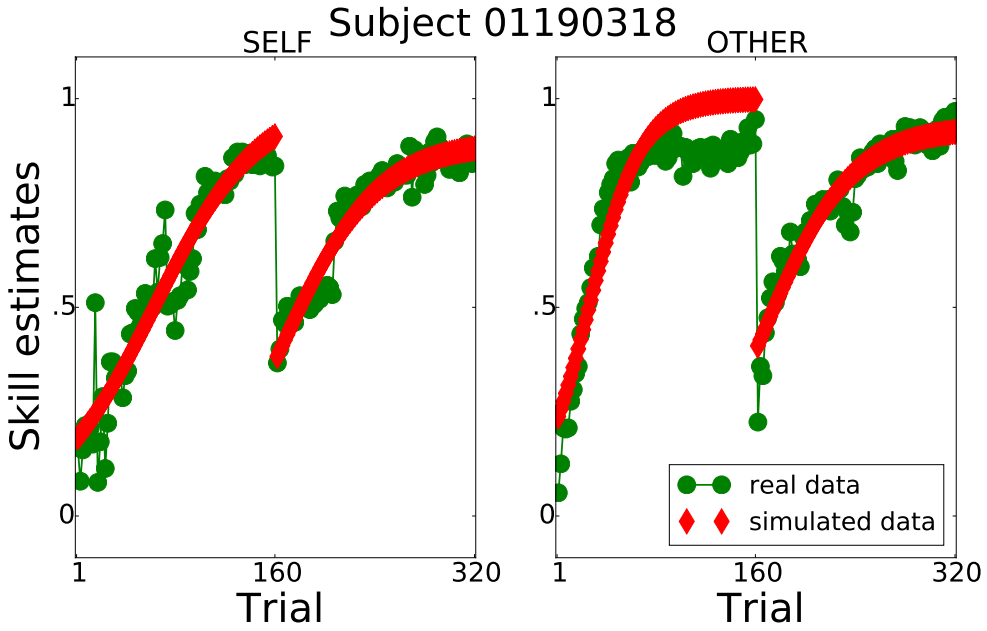


Figure 3.8: Example fit, model S3, subject with best fit. Green: real data, red: data generated from best fit parameters.

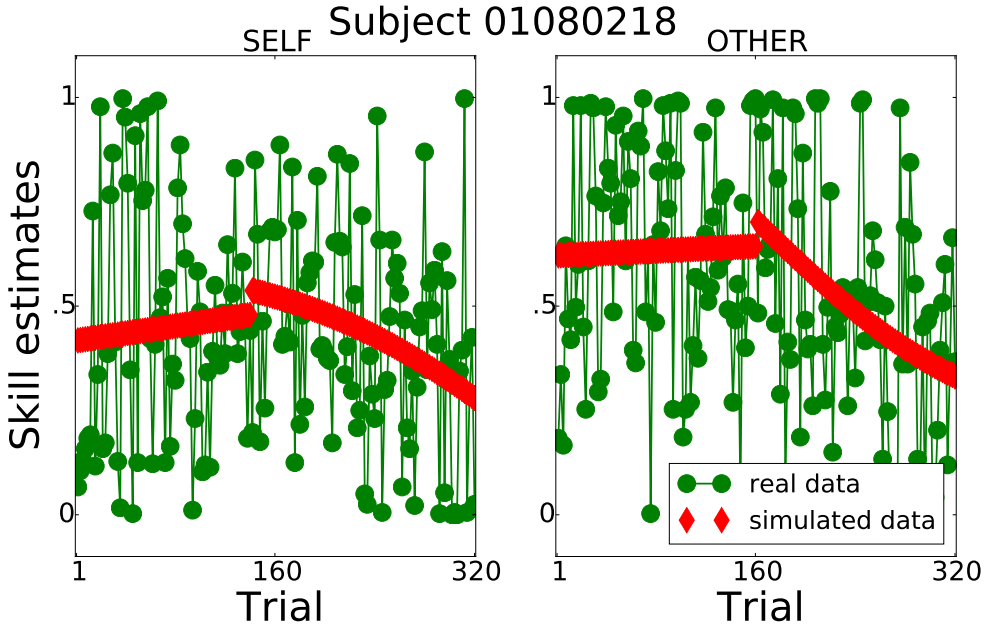


Figure 3.9: Example fit, model S3, subject with worst fit. Green: real data, red: data generated from best fit parameters.

3.3.2 Rescorla-Wagner models

Since we sought to investigate the mechanisms through which skill reports evolve, we next turned to models specifying these mechanisms.

One type of such models are Rescorla-Wagner(RW) models. Introduced (Rescorla, 1972) to explain learning curves for associability between stimuli and rewards, these models rely on the assumption that changes in associative strength are related to prediction errors - differences between the reward expected based on the current associative strength and the reward actually obtained. Models based on prediction error mechanisms have been immensely popular and successful in modelling animal and human behaviour in a wide variety of tasks and contexts (Siegel and Allan, 1996) and prediction error-like signals have been found in the activity of various brain regions, notably, but not only, related to dopamine signalling (Garrison et al., 2013), (see review in 1). In our task, rather than learning about reward, subjects can be assumed to learn about their own skill - a hidden variable- from experience at different levels of granularity, from key press to trial outcome.

All RW models were built on a common basic structure. Subjects' reported skill estimates,$\{r_t\}$, are assumed to be noisy readings of an evolving internal estimate of skill, $\{s_t\}$: $r_t \sim \mathcal{N}(s_t, \sigma)$[4]. On every trial, the prediction error δ_t between the experienced outcome o_t and the one expected from the current value of the internal skill estimate is computed, and used to update the internal skill estimate. This basic model involves two parameters: the initial value of the internal skill estimate, s_0, and the learning rate used to weigh the prediction error, α. Our baseline model, however, included an additional parameter modelling the effect of the break between sessions, β, resulting in the following model:

[4]All models were fitted separately for each subject in two ways: by minimising the sum of squared differences between model predictions and data (a least squares approach), and by fitting a full probabilistic version, including a noise response parameter. The list in J.2 contains the probabilistic versions. Results were similar between the two approaches, we present the least squares version in the following.

$$\delta_t = \begin{cases} o_t - s_{t-1} & \text{if } t \neq t_0^{II} \\ o_t - (s_{t-1} + \beta) & \text{otherwise} \end{cases}$$

$$s_t = \begin{cases} s_{t-1} + \alpha * \delta_t & \text{if } t \neq t_0^{II} \\ s_{t-1} + \beta + \alpha * \delta_t & \text{otherwise, where} \end{cases}$$

t_0^{II} = index of first trial of the second session.

Two mechanisms drove the expansion of the RW model family: we refined models based on assumptions about the factors of interest in order to test whether this enables them to better explain our data; and we built alternative models which ignored the factors of interest, in order to test to what extent their contribution is essential to explaining the data. Some of the variations between the RW models we compared were formalised as modulations of the learning rates by the factors of interest, or the introduction of separate parameters for the two sessions. Other models included changes to the outcome prediction in order to include difficulty information, or changes in the response, in order to include effects at different timescales.

In this subsection we present some of the RW models we compared, focusing on the factors that were responsible for various changes or additions in our models. See appendix J.2 for a complete list of the models, fully specified.

Baseline

We started from the baseline model defined above, assuming learning from outcomes, since this is the most salient source of information subjects had about their own evolving skill. The baseline also took into account the existence of the two sessions, by having a session break effect parameter. We also fitted a richer version of this baseline, allowing learning rates to be different between the two sessions - model `S___`[5].

We then fitted more complex models, which included the factors that we

[5]For most of the models presented in this section, the naming convention reflects the presence or absence of four orthogonal factors, labelled S,A,O,T (see this section's text for description). Other models, such as those including difficulty, follow slightly different naming conventions. All models are fully specified in appendix J.2

were interested in, such as attribution, difficulty and timescale of updating. There were large individual differences between the effects of these additions on the quality of fit, which we illustrate below.

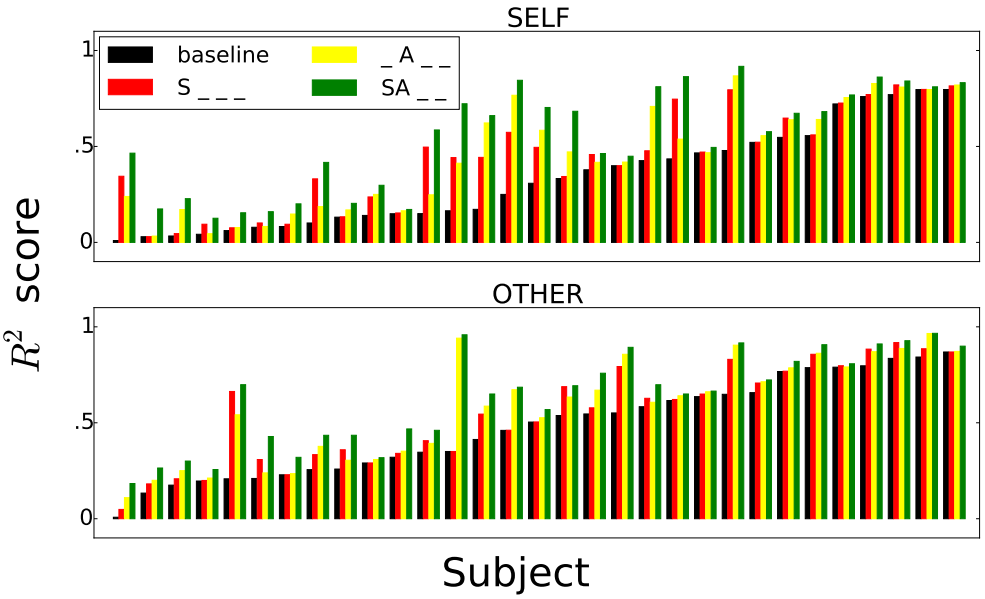


Figure 3.10: Effect of adding attribution to baseline models. Black: baseline model - one learning rate for both sessions, session break effect; red: baseline + different learning rates for the two sessions (A___); yellow: baseline + different learning rates according to attribution (_A__); green: baseline + different learning rates according to session and attribution (SA__). Individual subjects scores, ordered according to the baseline model scores, separately for 'self' and 'other'.

Baseline + attribution

First we augmented the baseline models by allowing learning rates to depend on attribution (resulting models are _A__ and SA__). See figure 3.10, which shows the r^2 scores of the best fit for the baseline models and their versions augmented with attribution information [6] Although in some cases adding attribution to the baseline models significantly increased the quality of fit, this was not a general pattern. Note that our way of augmenting models with attribution was by allowing different learning rates for outcomes attributed internally, externally and for outcomes lacking attributions; this latter parameter is one that we had to include due to our choice of only asking

[6]Note that this figure only shows the scores for the best fitting parameters, without any model complexity penalty, and we wouldn't use it as such for model comparison purposes. We will present model comparison analyses later (see 3.3.4), but use r^2 scores here since they are more easily interpretable.

subjects for an attribution every two trials. Thus an umbrella parameter was used for potentially different real underlying attributions, which can impair our ability to correctly identify the impact of attribution.

Baseline + attribution + timescale

We then considered the possibility of different timescales of belief updating. To model this, in addition to the learning mechanism represented by the prediction error learning, we augmented our attribution models by allowing outcomes to have an effect on the immediately following skill estimate (we refer to this as the outcome impulse), which does not propagate into future estimates (models `_A_T` and `SA_T`). This is both a common sense idea, and inspired by the response patterns of some of our subjects, who have high variation between successive responses. We suspected these might be due to a large effect of the just experienced outcome. Figure 3.11 shows the r^2 scores of these models compared with the baseline model and the enriched baseline model with attribution information, `SA___`.

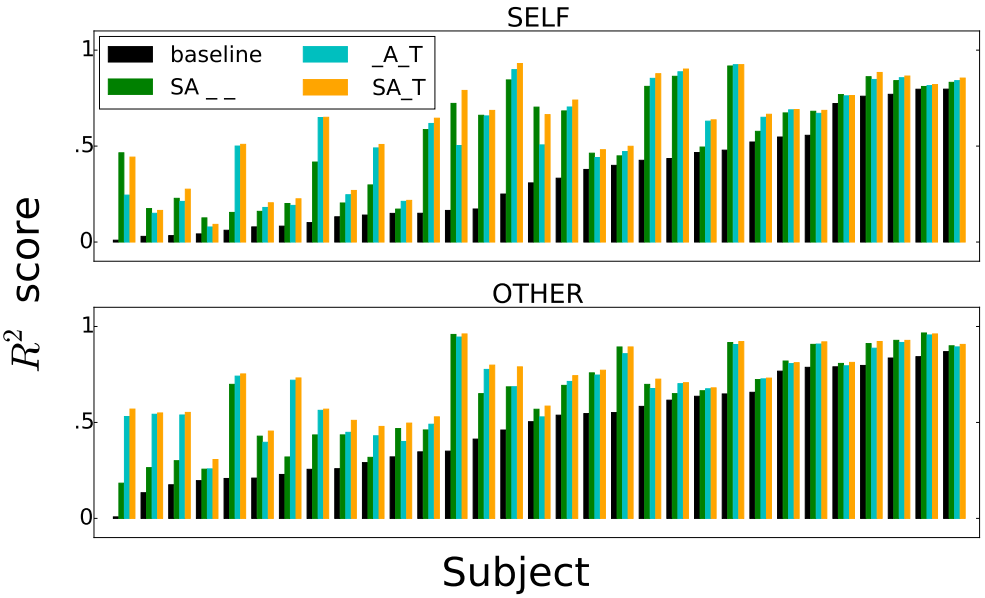


Figure 3.11: Effect of adding outcome impulse to attribution models. Black: baseline model; green: baseline + different learning rates depending on session and attribution (`SA___`); cyan: baseline + different learning rates depending on attribution + outcome impulse (`_A_T`); orange: baseline + different learning rates depending on session and attribution + outcome impulse (`SA_T`). Individual subjects scores, ordered according to the baseline model scores, separately for 'self' and 'other'.

Adding this local effect of outcomes seems to have a significant effect in improving the r^2 score for a very small number of subjects, notably ones with very low scores. However it seems to have little or no effect for subjects which are already somewhat better fitted. Note that here as well, no penalty has yet been applied for model complexity.

We checked whether the subjects for whom large improvements in r^2 scores result from adding outcome impulses are the ones with highly variable skill estimates, and this seems indeed to be the case. See figure 3.12 for an example subject.

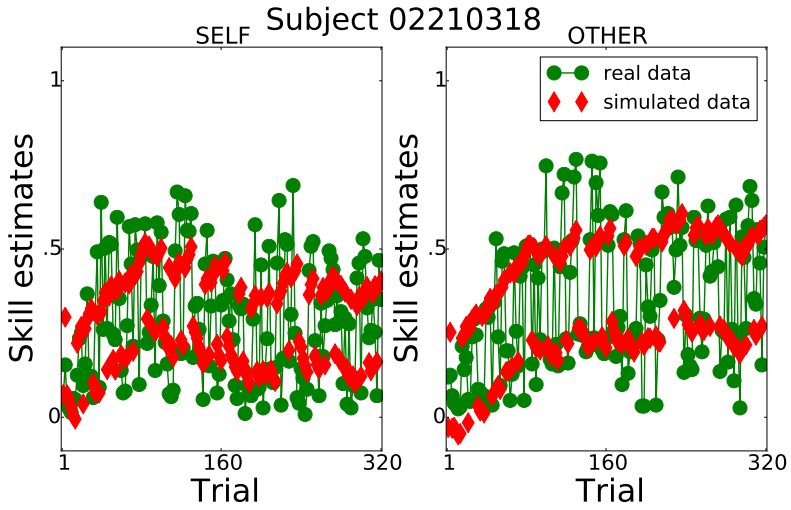


Figure 3.12: Example subject for whom adding the outcome impulse improves quality of fit: best fit for model _A_T

Because attribution on its own did not seem to significantly increase quality of fit for these subjects, and in order to test if the outcome impulse on its own is enough to produce this improvement, we also fitted baseline models with an outcome impulse, but without attribution dependency (models __T and S__T). Figure 3.13 shows the best fit r^2 for the baseline and for models ____T, _A___ and _A_T. Baseline + timescale

Two general observations arise from figure 3.13: first, although generally adding the outcome impulse parameters to the baseline model(resulting in ____T) increases its score, the further addition of attribution (model _A_T) has, in most cases, a much smaller effect (although there are exceptions, most

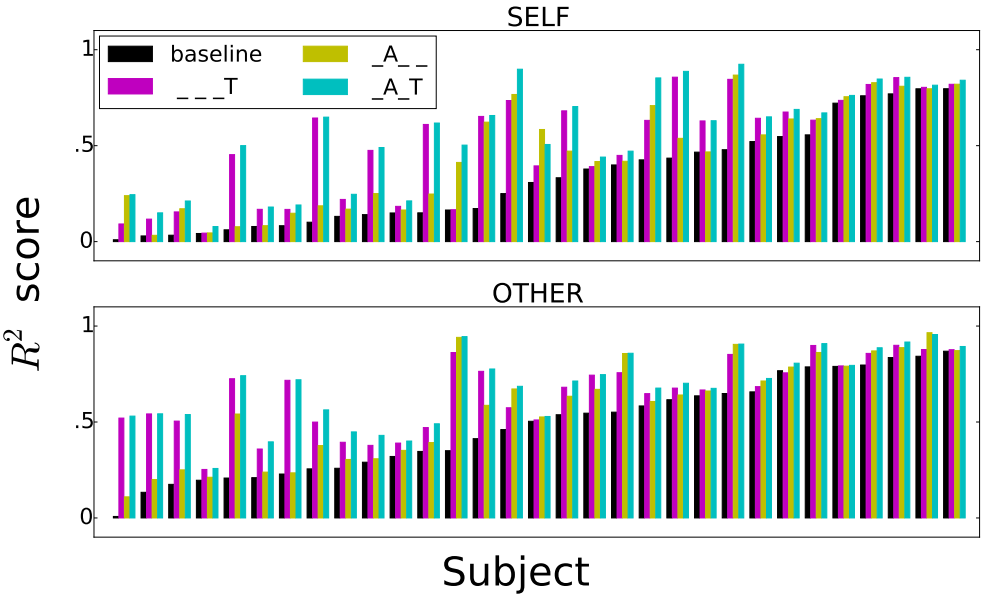


Figure 3.13: Effect of adding outcome impulse to baseline vs allowing learning rates to depend on attribution. Black: baseline model; purple: baseline + outcome impulse (____T); gold: baseline + different learning rates depending on attribution (_A___); cyan: baseline + different learning rates depending on attribution + outcome impulse (_A_T). Individual subjects scores, ordered according to the baseline model scores, separately for 'self' and 'other'.

notably the "self" condition). Secondly, model ____T scores better than _A___ in general, indicating that in most cases outcome impulse has more explanatory power over the baseline model than attribution.

We note, however, that about a third of the subjects in both conditions show the opposite pattern, again indicating variability in the mechanisms subjects might be using (note that these two models have the same number of parameters, and comparisons between their scores can meaningfully be made without worrying about penalising complexity).

Adding an outcome impulse improves model fits despite the fact that a mechanism for learning from outcomes is already included in the baseline model. There are two ways in which the outcome impulse models are richer than the baseline: they allow outcomes to influence skill estimates on two different timescales, and they allow positive and negative outcomes to impact learning differently. Indeed the fact that humans as well as animals learn dif-

ferently from positive vs negative outcomes has been extensively documented (Yacubian et al., 2006; Wrase et al., 2007; Frank et al., 2007; Seymour et al., 2007; Cools et al., 2008; Sharot et al., 2011; Cazé and Van Der Meer, 2013; Cox et al., 2015). This is consistent with some of our model agnostic analyses results, that show the effect of attributions to be different for wins and losses.

Baseline + attribution + outcome valence

We therefore again expanded our model family, by fitting models which, in addition to allowing attribution to modulate learning from outcomes, allowed positive and negative outcomes to differently modulate learning, or to modulate the effect of attribution (models _AO_ and SAO_). Figure 3.14 shows the best fit r^2 comparison between adding attribution, outcome impulse, and outcome-based learning rates to the baseline model.

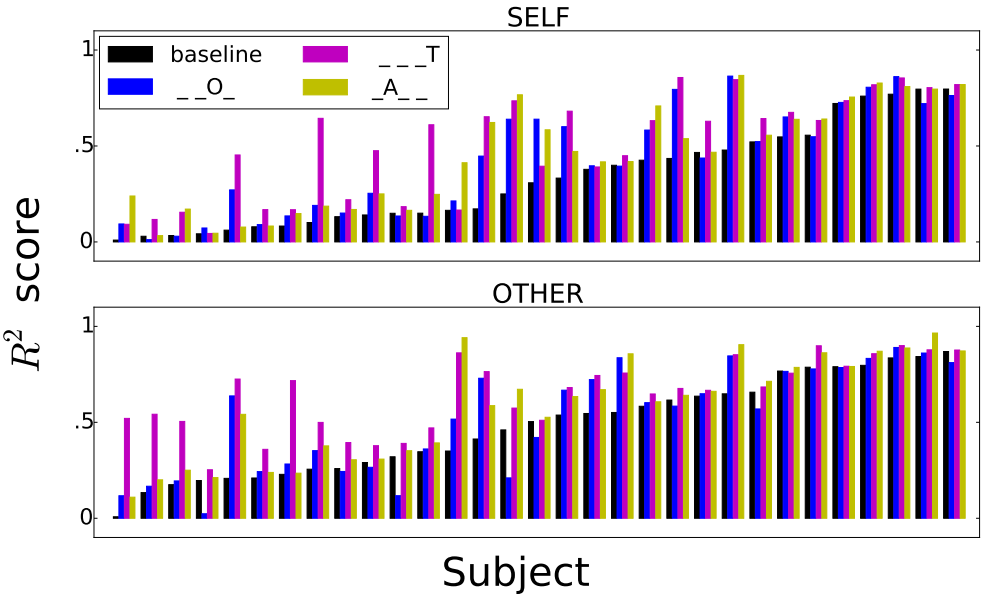


Figure 3.14: Effect of adding outcome impulse to baseline model vs allowing learning rates to depend on outcome vs allowing learning rates to depend on attribution. Black: baseline model; purple: baseline + outcome impulse (____T); gold: baseline + different learning rates depending on attribution (_A___); blue: baseline + different learning rates depending on outcome (__O_). Individual subjects scores, ordered according to the baseline model scores, separately for 'self' and 'other'.

Figure 3.14 shows that adding impulse response improves the baseline model more than simply allowing differentiated learning from positive and negative outcomes, on only one timescale. It also shows that in general allow-

ing different learning rates for different attributions does better than allowing different learning rates for wins and losses. Note, however, that as in all the previous r^2 comparison figures, there is no penalty for model complexity. We will move on to model comparison shortly.

Difficulty

We also built a number of models to examine whether difficulty modulates skill updating, for instance through changing learning rates on its own, over different timescales, or interacting with attribution. Detailed descriptions of these models can be found in appendix J.2). We do not illustrate them in this section, because in terms of best fit r^2 scores we observed similar patterns with the previous models: there are subjects who are poorly fit, and none of the models does much to improve this; there are subjects which are very well fit by the baseline model, and any improvement is marginal, and there are some subjects for whom accounting for difficulty does lead to significant improvement, as indeed is the case with other model augmentations

RW models summary

To summarize the model family we have presented in this section: there are four factors that our models can independently account for or not: session, outcome, timescale and attribution, resulting in 16 different models. Outcome and attribution are accounted for by allowing learning rates to differ according to the different values of these factors; timescale is encompassed by including outcome impulse parameters; session is addressed by allowing parameters to be different between sessions. In addition to these, we compared a number of models including effects of trial difficulty, described in appendix J.2. We introduced these factors into our models gradually, for different reasons: attribution was our original interest; session effects might have appeared as an artifact of our choices for practical experimental timing, outcome, difficulty and timescale are both theoretically interesting and factors that emerged as important in our model-agnostic analyses and early modelling work.

3.3.3 Observing the observer models

Our RW family ended up being quite extensive, but models were closely related, having the same underlying principle. We also fit models of a different family, observing the observer(OO) models (Daunizeau et al., 2010b,a), which we describe in this section. As we have already illustrated the factors driving the expansion of the model family in the case of the RW models, we will not go through the same process here; instead, we present the resulting models directly.

Observing the observer models rely on the following assumptions: subjects have a model of how the data they can observe is generated (perceptual model); they invert this model to compute the posterior belief on the underlying causes conditioned on the data they observed; they then choose a response based on their belief - this is the response model. Model comparisons between different perceptual and response models can be made within this framework.

In such models, the propagation of subjects' beliefs about the underlying variables generating the observed data is a filtering process (Bishop, 2006), which involves two steps for processing every trial. The first step is to turn a prior belief propagated from the end of the previous trial into a posterior belief by conditioning on the observations of the current trial. The second step is to propagate this posterior belief forward through a function encapsulating subjects' expectations about how these variables evolve in time (the link function). This produces the prior belief appropriate to the next trial. Apart from very particular cases, generally involving manipulations of Gaussian distributions, this process is not straightforward. This was also the case with our models, which involved both Gaussian and non-Gaussian distributions (see below). See appendix K for a detailed account of the computations involved.

To rigorously define the OO models we fit and to illustrate the differences

between them, the following notations will be used:

s_t : real, hidden value of skill at trial t

o_t : outcome of trial t

d_t : difficulty of trial t

r_t : skill estimate at trial t (subject's response)

$\mathcal{N}(\mu_{old,t}, \sigma^2_{old,t}) \approx p(s_t|o_1, ...o_{t-1})$: Gaussian approximation of subject's belief about skill before seeing trial t

$\mathcal{N}(\mu_{new,t}, \sigma^2_{new,t}) \approx p(s_t|o_1, ...o_t)$: Gaussian approximation of subject's belief about skill after seeing trial t .

The models that we compared were built on a common structure.

We assumed subjects' perceptual model consists in a skill evolution model, encapsulating their expectations about how skill evolves in time, and a generative model for the outcome, representing their beliefs about how skill and difficulty relate to outcome on a given trial. The skill evolution model we assumed to be a Markov time series, but we allowed the link function between the time steps to vary between different models (see below). In all cases the generative model for outcomes was the same, namely assumed to be a Bernoulli variable with the probability of winning dependent on the current difficulty and skill level.

$$o_t \sim Bernoulli(\sigma(s_t - d_t)).$$

The response model was also common to all models. We assumed subjects compute the probability of winning based on their current belief about skill and the difficulties they have recently encountered and respond with a

Gaussian noise-corrupted report of this value.:

$$r_t \sim \mathcal{N}(\sigma(\mu_{new,t} - \overline{d}_t), \sigma_r), \text{ where}$$
$$\mu_{new,t} = \text{ expected value of underlying skill, after trial t}$$
$$\overline{d}_t = \frac{d_t + d_{t-1} + ... + d_{t-10}}{10}$$

The models that we compared varied along three dimensions: subjects' internal expectations about how their skill would evolve in time (the link function in the perceptual model), belief updating mechanisms reflecting some of our factors of interest analysed with the RW models, namely attribution and outcome valence, and whether model parameters were allowed to differ between sessions[7].

The link function in the **skill evolution model** could be either a random walk (coded `R`), in which case

$$s_t \sim \mathcal{N}(s_{t-1}, \sigma^2_{process})$$

or an additive (potentially subtractive) linear model (coded `L`), in which case

$$s_t \sim \mathcal{N}(\alpha + s_{t-1}, \sigma^2_{process}).$$

The **belief updating** - the process by which an experienced outcome changes the distribution of beliefs over current skill - could be either an approximation to the normative Bayesian update (coded `B`)

$$\sigma^2_{new} = \frac{1}{\frac{1}{\sigma^2_{old}} + (1 - \sigma(\mu_{old} - d_t))\sigma(\mu_{old} - d_t)}$$
$$\mu_{new} = \begin{cases} \mu_{old} + (1 - \sigma(\mu_{old} - d_t))\sigma^2_{new} & \text{if } o_t = 1 \\ \mu_{old} - \sigma(\mu_{old} - d_t)\sigma^2_{new} & \text{otherwise,} \end{cases}$$

[7]The naming convention reflects the presence of these three orthogonal factors, model names being formed as `OO` followed by a three letter combination coding for session, belief updating and link function, respectively. See text for details.

or a modulation of it based on attribution (coded A) or outcome (coded O), meant to model any additional effect of attribution or outcome over and above the normative update(the detailed derivation of these is provided in appendix K):

$$\sigma^2_{new} = \frac{1}{\frac{1}{\sigma^2_{old}} + \alpha(1-\sigma(\mu_{old}-d_t))\sigma(\mu_{old}-d_t)}$$

$$\mu_{new} = \begin{cases} \mu_{old} + \alpha * \frac{(1-\sigma(\mu_{old}-d_t))}{\frac{1}{\sigma^2_{old}}+(1-\sigma(\mu_{old}-d_t))\sigma(\mu_{old}-d_t)} & \text{if } o_t = 1 \\ \mu_{old} - \alpha * \frac{\sigma(\mu_{old}-d_t)}{\frac{1}{\sigma^2_{old}}+(1-\sigma(\mu_{old}-d_t))\sigma(\mu_{old}-d_t)} & \text{otherwise, where} \end{cases}$$

α varies according to attribution or according to outcome.

Finally, the two sessions could either share parameters (coded _) or were allowed to have different parameters (coded S). In all cases an effect of session break was included as follows: for the first trial of the second session in all the updates σ^2_{old} and μ_{old} were replaced by $\sigma^2_{old} + \sigma^2_{break}$ and $\mu_{old} + \mu_{break}$ respectively.

The combination of these different model choice options resulted in 12 models, listed in table 3.2.

	Random walk link		Additive link	
	Same params	Different params	Same params	Different params
Bayesian update	OO_BR	OOSBR	OO_BL	OOSBL
Attribution modulation	OO_AR	OOSAR	OO_AL	OOSAL
Outcome modulation	OO_OR	OOSOR	OO_OL	OOSOL

Table 3.2: OO models. See text for description.

3.3.4 Model comparison

We used BIC scores for model comparison, see appendix J.3 for details related to the BIC scores computation.

In general our RW and OO models did much better than the purely descriptive ones, as expected, indicating that the mechanisms we postulated in-

deed helped explain subject's responses. Comparison of BIC scores summed over subjects showed that for both "self" and "other", none of the purely descriptive models was among the top 10 models, with ΔBIC$=643.74$ between the best model and the best descriptive model for "self" and ΔBIC$=697.6$ for "other". This pattern was also present at the individual subject level, model comparison preferring one of the RW or OO models to the purely descriptive models with a $\Delta\text{BIC} \geq 10$ for 18 out of our 31 subjects in the "self" condition, and for 20 out of the 31 in the "other" condition (see also figure 3.7).

For both conditions, the winning model according to the summed BIC comparison was a RW model (see figure 3.15). For "self" the preferred model was `S_OT` - including different learning rates for wins and losses, outcome impulse effect and different parameters for the two sessions, but no effect of attribution - with a difference of 111.11 in BIC score w.r.t the second best model. For "other", the winning model was model `SA_T` - including different learning rates for different attributions, outcome impulse effect and different parameters for the two sessions- with a difference of 72.25 in BIC score w.r.t the second best model.

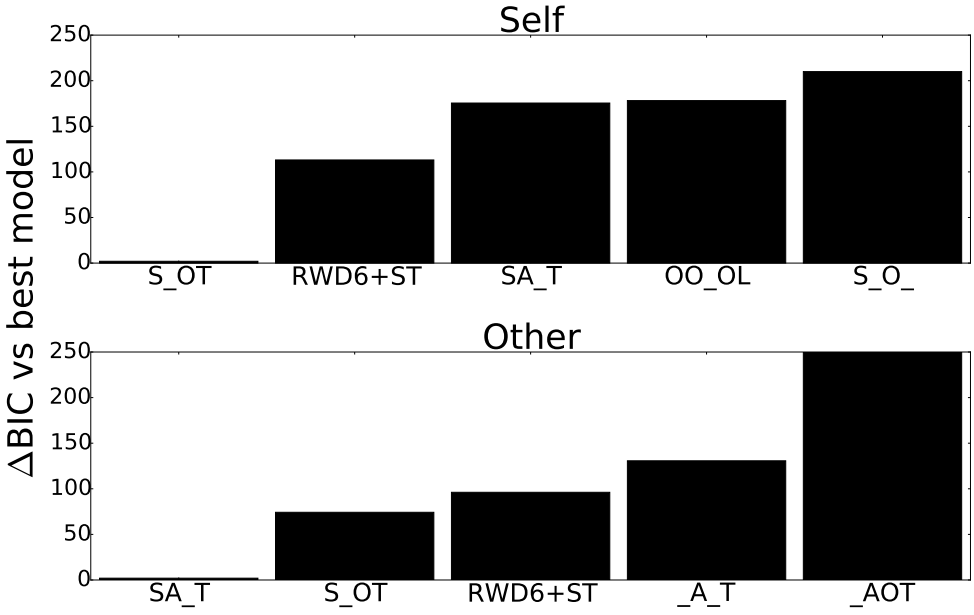


Figure 3.15: Model comparison results: summed BIC score, top 5 models.

However establishing the importance of the different factors of interest

is more complicated that it would appear from the summed BIC comparison: the winning model for the "self" condition at the population level was the winning model at the individual level for only 1 subject, and was among the top 5 for for only 3 subjects; for the "other" condition, the winning model at the population level was not the winning model for any of the subjects, but was among the top 5 models for 6 of the subjects.

As mentioned before, we found large variability between subjects, beyond that apparent in their patterns of skill estimates. Thus apart from variation in the extent to which our models managed to capture the data, there was also variability in the models that were preferred for different subjects and in the confidence with which they were preferred. For the "self" condition there were only 6 out of 31 subjects for whom the BIC score difference between the best and second best model was larger than 6 -indicating strong evidence in favour of the winning models- and no subject showed a difference larger than 10, indicating very strong evidence in favour of the best model. In the "other" condition only 5 subjects had a difference larger than 6 between the BIC scores for the best and second best models, and 2 of them had a difference larger than 10 between the two best models. In the "self" condition, 21 of the models that we compared (there were 47 in total) were the winning model for at least one subject, and in the "other" condition, the number of distinct models being the best model for at least one subject was 23.

These observations indicate that across subjects model comparison would not be adequate; however our data are also not sufficient for performing subject clustering. The fact that we ended up with a large number of models, varying along several independent dimensions (defined by the factors that models took into account), also poses its problems, as it can become difficult to interpret model comparisons if a clear pattern does not emerge across subjects, which seems to be our case. We do not therefore attempt to draw any general conclusion about the belief updating mechanisms at the level of the population, or about the existence and nature of different subgroups within it.

Instead, we performed further analyses at the individual subject level, aimed at determining whether any of the subjects showed significant effects of the factors of interest (attribution, difficulty, outcome impulse, outcome valence). As OO models did not generally do better than RW models, and comparisons between RW models can be performed without any additional complications introduced by structural differences, we limited these analyses to the RW models only[8]. We present these analyses in the remaining part of this section.

For this purpose we implemented the following pipeline of selection criteria, such that we were satisfied that any subjects who passed through the entire pipeline showed a convincing effect of the factor of interest: we first selected subjects for whom at least some models achieved a best fit r^2 score of at least 0.6; out of these, we selected the subjects for whom the best model included the factor of interest; we then selected the ones for whom the BIC score difference between the best model and the best model *without* the factor of interest was larger than 10. We then applied two more tests. The first was a permutation-test -like analysis, in which we randomly chose 5000 permutations of the values of the factor of interest, and for each permutation refitted the winning model and computed the BIC score. Our aim was to test if the BIC score obtained for the real ordering of the values of the factor of interest was different from the distribution of BIC scores obtained by permuting these values. We rejected subjects for whom the approximated p-value was larger than 0.05. Finally, for all remaining subjects we tested model identifiability, by generating data from the best fit parameters of the best fitting model and fitting both the best model and the best model without the factor of interest to

[8]We have, however, checked wether any subjects showed strong preference for the OO models vs the RW ones (ΔBIC $\geq$ 10 in favour of an OO model) and found 5 such subjects for "self" and one for "other". We note that none of these were among the subjects for whom we found significant effect of our factors of interest in the RW-only analyses(see below). This suggests that subjects might differ with respect to the underlying mechanism of belief updating, and not only with respect to the factors influencing learning in a RW context. Further investigation of this additional source of variability in the subject population remains a goal for future work.

this simulated data. Subjects for whom the correct model was not preferred would be rejected.

We present below the results of applying these criteria for each factor of interest in turn.

There were 19 subjects with a best r^2 score larger than 0.6 for "self", and 23 satisfying this criterion for "other"; these were the subjects included in the analyses below.

Attribution: Applying the selection criteria above for attribution we found 6 subjects for whom the winning model included attribution in the "self" condition and 10 in the "other" condition. Out of these, in each condition two had a $\Delta\text{BIC} > 10$ between the best model and best model without attribution. Both subjects for the "other" condition survived the attributions permutation test and the model identifiability test. Only one of them survived in the "self" condition, the other failed the attribution permutation test. We note that in the "self" condition there was an additional subject for whom the BIC score difference between the best model without attribution and the best model was 9.98. We performed the attributions permutation test and the model identifiability test for this subject as well and found that the subject would have survived these.

Difficulty: Applying the selection criteria above for difficulty we found that for both conditions 8 subjects had a winning model including difficulty. Out of these, three had a $\Delta\text{BIC} > 10$ between the best model and best model without difficulty and all three survived the additional permutation and model identifiability tests for the "self" condition. Only one subject in the "other" condition passed the ΔBIC threshold, but failed the difficulty permutation test. We note that we performed the difficulty permutation and model identifiability tests for two additional subjects that did not pass the ΔBIC threshold, but had ΔBIC values of 9.6 and 9.5; one of these failed the difficulty permutation test, but the other survived both additional tests.

Outcome impulse: Applying the selection criteria above for outcome

impulse we found that 13 subjects had a winning model including outcome impulse in the "self" and 11 in the "other" condition. Out of these, 8 passed the ΔBIC selection criteria for "self" and 5 for "other". All of these survived the impulse permutation and model identifiability tests.

Outcome valence: Finally, for outcome valence there were 6 subjects with a winning model including outcome valence in the "self" and 5 in the "other" condition. Three of those in the "self" and two of those in the "other" condition passed the ΔBIC criterion. They all survived the following two tests.

We found therefore that for each of the factors of interest there were subjects for whom these factors were an important part of the belief updating mechanism. The precise values of our thresholds in the pipeline were somewhat arbitrary, but they were chosen so as to impose harsh thresholds in the pipeline, and therefore assure that subjects surviving all tests show convincing evidence in favour of the respective factors.

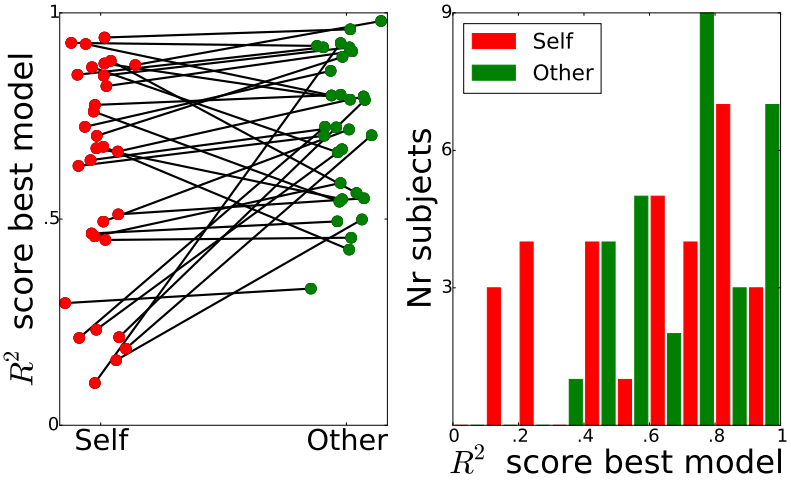


Figure 3.16: Distribution of cross-validation r^2 scores for best models, both conditions. Left: score of best model for 'self' vs 'other'. Right: distribution of highest scores across subjects, counts.

Finally, as far as the comparison between the two conditions is concerned, data from the "other" condition was in general better captured by our models than data from "self" (cross-validation r^2 scores of best models for "self" vs "other": paired $t(15) = -2.32, p = 0.027$); see figure 3.16.

3.3.5 Model-dependent analyses: summary

In this section we presented model dependent analyses of subjects' skill estimates. We compared a large number of models: purely descriptive models, models based on prediction error updates (RW models) and models involving more complex accounts of subjects' internal beliefs and their evolution (OO models).

Contrary to our expectations, rather than uncovering common patterns underlying the evolution of subjects' skill estimates, these revealed a large amount of variability between subjects, both in the extent to which our models could explain the data, and in the models which best did so. As a result, and due to the size of our dataset, making any general inferences about the belief updating mechanisms at the level of the population or about the existence and nature of different subgroups within it was not possible.

Instead, we performed further analyses at the individual subject level, aimed at determining whether any of the subjects showed significant preference for models including each of the factors of interest (which, according to model-agnostic analyses, had significant effects on skill estimates at the population level). This was checked by testing a series of conditions in a pipeline, designed such that we could be satisfied that any subject resulting from this selection process displayed convincing evidence of the use of the respective factor of interest. The results of these further analyses showed that there were indeed subjects whose responses were better explained by models including difficulty information, subjects for whom learning from outcomes was modulated by their reported attribution of the outcome, subjects who learned differently from wins vs losses, and subjects whose responses reflected short term effects of outcome.

Models were generally better at explaining data from the "other" condition than data from the "self" condition, which is consistent with model-agnostic analyses, and suggests responses provided for "self" are noisier; this might be due to the heightened relevance of the "self" condition, to differ-

ences between acting and watching, or to the ordering of the two conditions.

These model-dependent analyses highlighted the variability between subjects, emphasising the need for larger data sets that could be used to investigate potential clusters of subjects with common mechanisms. They also showed that effects of the different factors of interest, and in particular attribution, can be detected at the individual subject level.

3.4 Reaction times

In this section we present analyses of reaction times (RTs) for reporting skill estimates, and their relationships with outcome, attribution, reported skill and condition.

In addition to the deliberation associated with skill reporting, which is what we are interested in investigating, RTs also reflect inter-individual and trial by trial variability in movement speed and trial to trial variability in the (randomly chosen) initial slider position(see 2). To control for these, we distinguished three components of the raw reaction time: the time until the first key press, which marks the start of the response - we refer to this in the following as the "latency"; the total time of the slider movement; and the time to submit the response by pressing the ENTER key -we refer to this as the "submission time".

The variables we analysed were the residual RTs, defined, for each trial of a given subject, as the difference between the raw reaction time and the sum of the subject's median latency, the subject's median submission time and the time it would take to move the slider from the initial to the final position by holding the correct arrow key pressed:

$$rt_t^s = RT_t^s - (l^s + st^s + f(\Delta x_t^s)), \text{where}$$

rt_t^s = residual RT at trial t for subject s

RT_t^s = raw RT at trial t for subject s

l^s = median latency for subject s

st^s = median submission time for subject s

$f(\Delta x_t^s)$ = time to move the slider from the initial to the final position by holding correct key pressed.

For each of the factors of interest, (outcome, attribution, condition, skill estimates) we computed for each subject the average residual reaction time for all trials corresponding to the relevant factor level (we used quartile discretization for skill estimates) and compared the resulting distributions across subjects. Skill estimates and reaction times were z-scored within subject.

Overall, we found that subjects were significantly faster in responding in the 'self' vs the 'other' condition (paired $t(15) = -3.24, p < 0.01$, Hedges corrected d= 0.57). Such effects have been reported before (Jackson et al., 2006; Kuiper and Rogers, 1979; Nowicka et al., 2018) and might be due to subjects being more engaged when they are playing, compared to when they are watching; alternatively, it could be due to higher uncertainty in the 'other' condition, due to the lack of direct experience of the movements, or to concerns coming into play when reporting evaluations of others (Crockett et al., 2014; Rand et al., 2014), (see Rand and Nowak, 2013, for a review).

We found no significant difference in reaction times for responding after wins vs losses in either condition ("self": paired $t(15) = -1.65, p = 0.11$, d= 0.58; "other": paired $t(15) = 0.03, p = 0.98$, d = 0.01).

We found that subjects were significantly faster in responding after outcomes attributed internally vs externally in both conditions ("self": paired $t(15) = -3.22, p < 0.01$, d = 1.07; "other": paired $t(15) = -2.39, p = 0.02$,

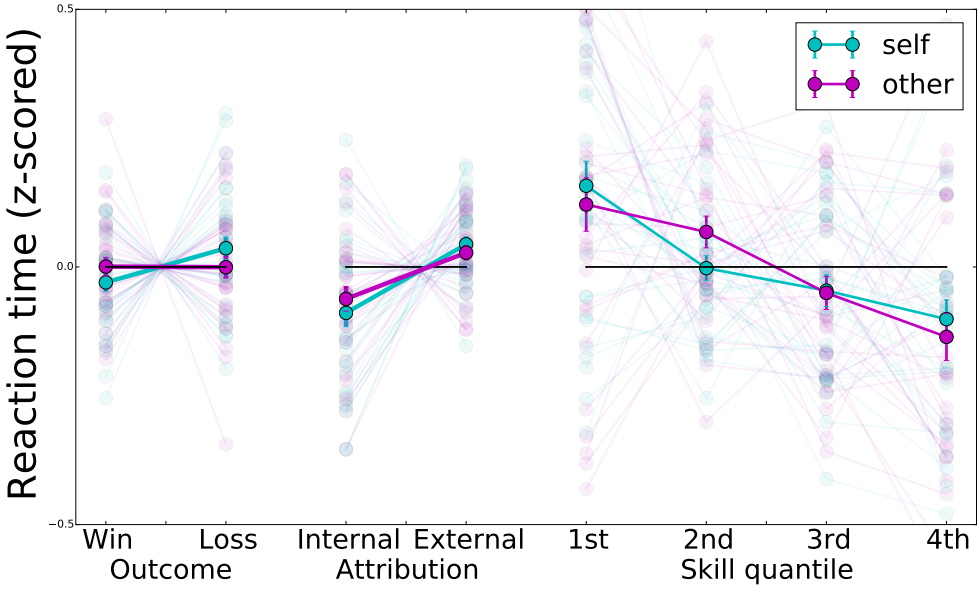


Figure 3.17: Effects of outcome, attribution and skill estimate on RT, mean ± s.e.m across subjects. Left: effect of outcome on RT. Middle: effect of attribution (internal vs external) on RT. Right: effect of the provided skill estimate on RT.

d = 0.8). This might reflect more engagement when assigning responsibility to oneself or another person, versus to more neutral circumstances, or increased complexity in computing what the trials outcome implies for the skill evolution when the outcome is perceived as being mainly due to external circumstances.

We performed repeated measures 1-way ANOVA, with skill report as a fixed factor and subjects as random factors to test whether there is any effect of the skill report provided on the reaction time. We found a significant effect in both conditions ("slef": $F_{(3,90)} = 6.91, p < 0.01$; "other": $F_{(3,90)} = 5.76, p < 0.01$), reaction times decreasing with increasing skill estimates (see figure 3.17). Thus subjects are faster when providing higher estimates of skill, for both "self" and "other", which is consistent with a preference toward positive evaluation of both the self and others. Note, however, that skill estimates generally increased across trials, and therefore time and skill estimate are confounded.

3.5 Summary

In this chapter we presented analyses of skill estimates, as well as the corresponding reaction times, investigating the effects of outcome, attribution response, difficulty and performance on the evolution of skill estimates in the two conditions.

Model agnostic analyses revealed, in both conditions, a significant effect of the latest outcome on the change in reported skill estimates, in the expected direction - higher increses after wins vs losses - as well as the presence of a recency effect, with outcomes two trials back having a smaller effect on skill updates than the immediately preceding outcome.

We also found significant effects of difficulty on skill updates after wins, but not losses, and significant effects of several measures of performance on skill updates after losses, but not wins, for both conditions.

Finally attribution - the factor of particular interest to us - had no main effect on skill updates, but significant interactions with outcome and secondary effects in both conditions, in the expected directions: losses attributed externally lead to smaller decreases in skill reports than losses attributed internally, with the opposite pattern being present for wins. Surprisingly, the effects of attribution were stronger for losses than for wins in both conditions and appeared consistently stronger for "other" than for "self", although direct comparisons between conditions did not identify differences between conditions as statistically significant.

We then presented our model-dependent analyses. We compared a large number of models, fitted to individual data in its original time-series structure, and aimed at allowing a more detailed investigation of the mechanisms of trial by trial skill updating.

We expected model comparisons to produce evidence favouring one or a small number of models at the population level, which would have allowed us to draw general conclusions about the mechanisms subjects use, as well as make comparisons between conditions based on parameter estimates for

these putative winning models. This, however, is not what we found. Model-dependent analyses highlighted the presence of large between-subject variability, indicating the importance of obtaining that larger data sets, which would make clustering analyses possible. Explanatory models did performed better than purely descriptive ones, however there was a high level of individual variability manifest at different levels: in the extent to which our models were able to fit the data, in the models preferred, in the confidence of these preferences and in the pattern of presence or absence of factors of interest in well-performing models.

Therefore instead of performing model comparison and investigating the effects of interest at the population level, we turned to the individual level and applied a series of stringent selection criteria to establish wether any or our subjects convincingly displayed any of the effects of interest. We found that this was indeed the case for all the factors of interest, namely difficulty, attribution, different learning from positive and negative outcomes and momentary, as well as longer-term effect of outcomes.

Finally, analyses of reaction times revealed that subjects were significantly faster in responding in the "self" vs the "other" condition - a previously documented effect (Jackson et al., 2006; Kuiper and Rogers, 1979; Nowicka et al., 2018), as well as faster when providing higher vs lower estimates of skill in both conditions. There was no significant difference in reaction times after wins vs losses, but, surprisingly, subjects were significantly faster in reacting after trials attributed internally than after trials attributed externally, in both conditions.

3.6 Discussion

Our analyses showed that an effect of attribution on belief updating is detectable in our task and that it is consistent with our expectations. In addition, attribution seems to have an effect on reaction times for providing skill estimates, which we had not expected.

Our analyses also revealed unexpected differences between the effect that attributions have, conditioned on outcome. Different learning from different outcome valences has been extensively documented (Yacubian et al., 2006; Wrase et al., 2007; Frank et al., 2007; Sharot et al., 2011; Cazé and Van Der Meer, 2013; Cox et al., 2015), and indeed we found evidence of it in our model-dependent analyses. Our observations about the differential effect of attributions is consistent with the hypothesis that different valences might also influence learning through different effects of causal attributions on belief updates: under asymmetric weighting of losses vs gains, it might be advantageous, when improving one's model of the world, to focus on learning about one's effectiveness in situations resulting in negative outcomes, and less so in positive outcome situations. From this perspective, the fact that the same pattern holds when observing others is puzzling.

A number of interesting questions spring from these observations, providing hypotheses for future work. One such question relates to the extent to which the overall environmental rate of reward could skew these effects: would safer and richer environments push subjects' belief updates to be more accurately related to causal attributions? Alternatively, in cases where subjects could not benefit from improving their model of the world, or in situations where they could benefit only from improving their knowledge of the positive outcome situations would these differences still be present? Furthermore, to what extent does the general level of control that they have or perceive to have over their environment modulate these effects, if at all? All these questions could be tackled by relatively simple environment manipulations (see 6.3).

Direct comparison between the effect of attribution in the "self" and "other" conditions did not identify the differences as significant, however effects of attribution were consistently stronger for "other" than for "self". If that is indeed the case, there are a number of factors that could explain it. Belief updating processes in the "self" condition might be more complex that

processes in the "other" condition, and people might be more honest when reporting attributions and or skill estimates for "other" than they are when reporting on their own performance, which is presumably more emotionally salient and engaging. However, these apparent differences might be merely a spurious effect, due to the presence of more noise in the "self" condition.

Alternatively, the difference between acting and watching might be entirely responsible for observed differences between the "self" and "other" conditions: in the "other" condition, subjects need only watch and evaluate , while they need to also act, and do so under time pressure, in the "self" condition.

Further investigation is needed to establish whether attribution has different effects when the self vs other people are involved. One simple manipulation likely to shed some light on this matter would involve subjects evaluating more external "others", for instance a real other and the fake other we used in this case, or a real other and a computer. Manipulating the in-group vs out-group affiliation of the "other", or the subject's emotional connection with them would also be useful in investigating this aspect's contribution to any self-other differences.

The role of difficulty and performance in driving skill updates also needs to be further investigated. Since this was the first implementation of the task, we did not have previously validated measures of performance and difficulty. Our simulation analyses (see 2.3.1) indicate that difficulty recovery is possible using our approach; however further testing and external validation of the difficulty measure we extracted is still needed, as is better calibration of the staircase mechanism. Asking subjects to rate difficulty and performance themselves would be a useful addition. It would allow comparisons with objective measures to be performed; it would also enable us to investigate relationships between estimated difficulty and performance and skill updates.

We found that performance has a stronger effect on skill updates following losses. This difference between wins and losses might reflect an adaptive

mechanism involving higher discrimination in learning from losses than from wins. It is consistent with a focus on learning from negative outcomes, in order to avoid them in the future, while allocating less discriminative attention to positive outcomes, which can more generally be used as a positive signal. In other words, it could be a sensible strategy to learn **that** something is good and focus on learning **why** or **how** something is bad. The fact that the effect of performance on skill updates after losses seems to be stronger for "self" than for "other" is consistent with this view, however we cannot with the present data establish whether differences between "self" and "other" are not merely due to differences in the perception of performance when playing vs watching. Although not significantly different, effects of difficulty appeared instead stronger for wins. This surprising observation hints at potentially different mechanisms for the integration of these two sources of information into the updating of beliefs, which future work could investigate.

The aim of our study was to propose a task for quantitatively investigating the postulated dynamical interaction between causal attributions and beliefs (Bentall et al., 2001; Bentall, 2003), which can be conceived as a loop involving reciprocal effects of the two variables. Being able to identify and investigate separately each of the two arrows constituting this loop is a necessary first step towards understanding the complexities of the system. The focus of our analyses in this chapter has been to establish whether our task is suitable for the identification and investigation of one of these arrows, namely the effect of attribution on belief updating. We conclude that this effect can indeed be identified in our task. The next chapter is dedicated to the effect in the opposite direction.

Chapter 4

Attributions

Causal attributions and beliefs about skill are the two main variables of interest in this study. In this chapter we present subjects' attribution responses and our analyses of these data.

As we did in the previous chapter, we start by presenting a summary of the data, focusing on the aspects that will be of interest in following analyses, namely subjects' preferences for the available response options, their evolution in time and the differences between responses in the "self" and "other" conditions.

We present model agnostic analyses of these data, focusing on several factors of interest: outcome, objective task and performance measures, subjects' skill estimates. We then move on to modelling and present model-dependent analyses of these factors. Finally we also present analyses of reaction times for the attribution responses.

The chapter ends with a discussion including our conclusions and directions for future work.

4.1 Data summary

Figure 4.1 shows all subject's responses to the attribution questions, for both conditions. As can be seen from figure 4.1 there is trial-to-trial variability in subjects' responses, and they use all the available options, although not uniformly. Some subjects show a clear preference for some options, such

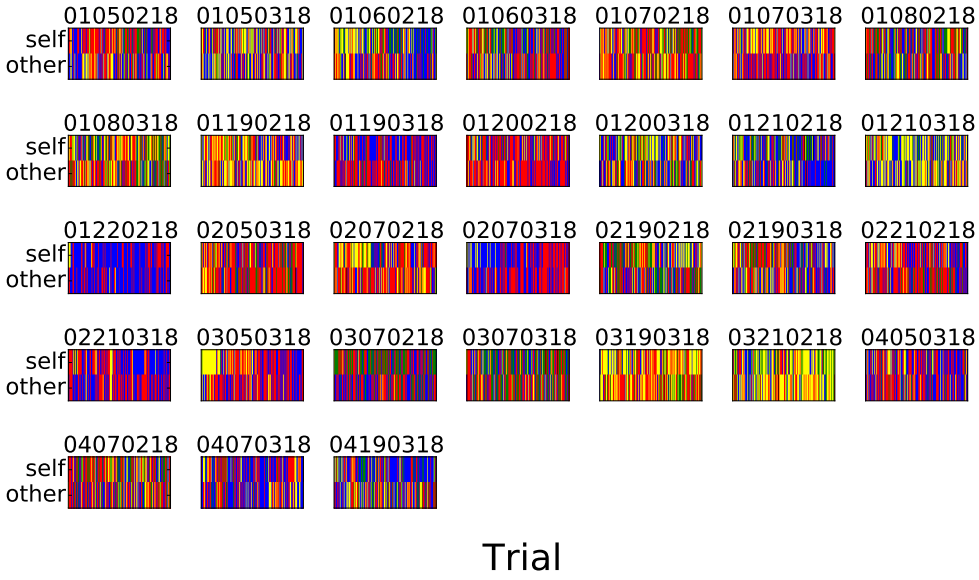


Figure 4.1: All attributions, all subjects. Each subject's attributions in each condition are represented by a series of vertical coloured strips. Each vertical strip represents the attribution response for a trial, color-coded according to the attribution option chosen: red - internal attribution, blue - attribution to maze, yellow - attribution to rotations, green - attribution to luck, black - missing attribution. Responses for "self" are plotted directly above responses for "other", for comparison.

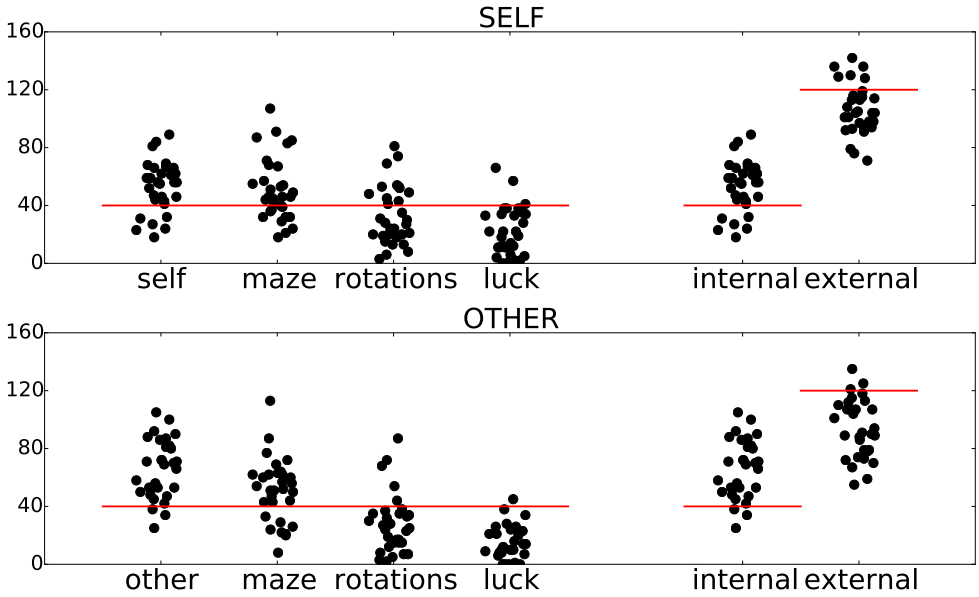


Figure 4.2: Total number of attributions for each option. Each dot represents a subject, the x-positions have been jittered for visualisation purposes. The red line indicates the number of attributions we would have seen, had subjects divided their attribution responses equally between the 4 options. Left: attribution counts for the 4 options. Right: attribution counts for relabelling the 4 options as 'internal' vs 'external'.

as subjects 01220218 (row 3, column 1) and 01190318 (row 2, column 3), who attributed outcomes mostly internally (red) and to the maze (blue), while others showed more mixed patterns, such as 01050318 (row 1, column 2) and 01080318 (row 2, column 1). We also note that no systematic difference between the two conditions is apparent.

Figure 4.2 shows the total number of attributions for each option, against the number expected by chance. In both conditions, most subjects provided more internal attributions (to "self" in the "self" condition, to "other" in the "other" condition) and attributions to "maze", and fewer attributions to "rotations" and "luck" that expected from a uniform distribution over the 4 options.

One concern was whether we had (inadvertently) introduced an availability bias in subjects' attribution responses, due to the fact that they were provided with three external attribution options and only one related to themselves (or the "other"). Figure 4.2 also shows the overall number of choices for "internal" vs "external" attributions obtained by relabelling the four available options accordingly. While subjects did indeed choose the three external options more than they chose the one internal option, they displayed a strong preference for making internal attributions, countering to some extent the availability bias, especially in the "other" condition.

Direct comparison between the two conditions shows that subjects made significantly more internal attributions for "other" than they did for "self" (paired $t(15) = -5.18$, $p = 10^{-5}$), consistent with the "actor-observer effect" (Jones and Nisbett, 1987) (see discussion in section 4.6).

Figure 4.3 shows the evolution over time of the number of attributions for each option. We note that while the average number of attributions for "self" / "other" is relatively constant in time, there is a slight increase, with time, in the average number of attributions to "maze" and "luck" and a slight decrease in the average number of attributions to "rotations". This is consistent with a scenario in which rotations' influence on subjects's performance decreases as they learn the task better, and in which subjects detect this change and respond

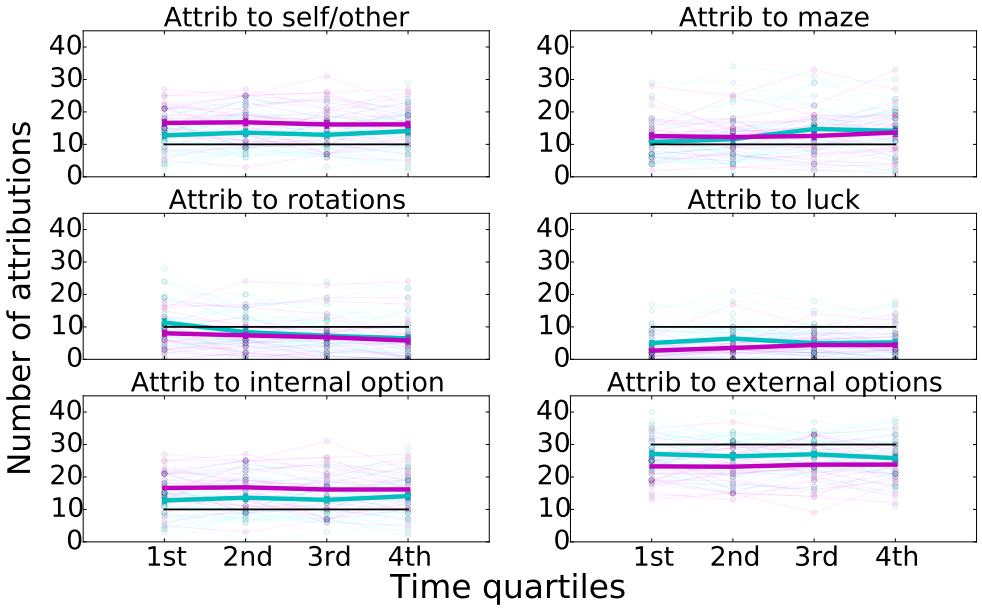


Figure 4.3: Total number of attributions for each option (top and middle), and for attributions relabelled as 'internal' vs 'external' (bottom), across four equal time periods in the task. "Self": cyan, "other": magenta. Faded lines represent individual subjects, thick lines represent averages ± s.e.m. across subjects. Black lines indicate the number of attributions we would have seen, had subjects divided their attribution responses equally between the 4 available options.

accordingly.

A final aspect of the data that we summarize in this section is the extent to which attribution responses differ between the two conditions, on a trial-by-trial basis. The top plot in figure 4.4 shows the proportion of mismatched attributions between the two conditions, considering all 4 options (black), as well as their relabelling as "internal" vs "external" (yellow). In both cases the amount of difference between attribution responses in the two conditions is on average relatively stable across time. Although the proportion of mismatches is quite high, subjects show less differences between responses in the two conditions than would be expected if their responses in the two conditions were independent (see the bottom plots in figure 4.4).

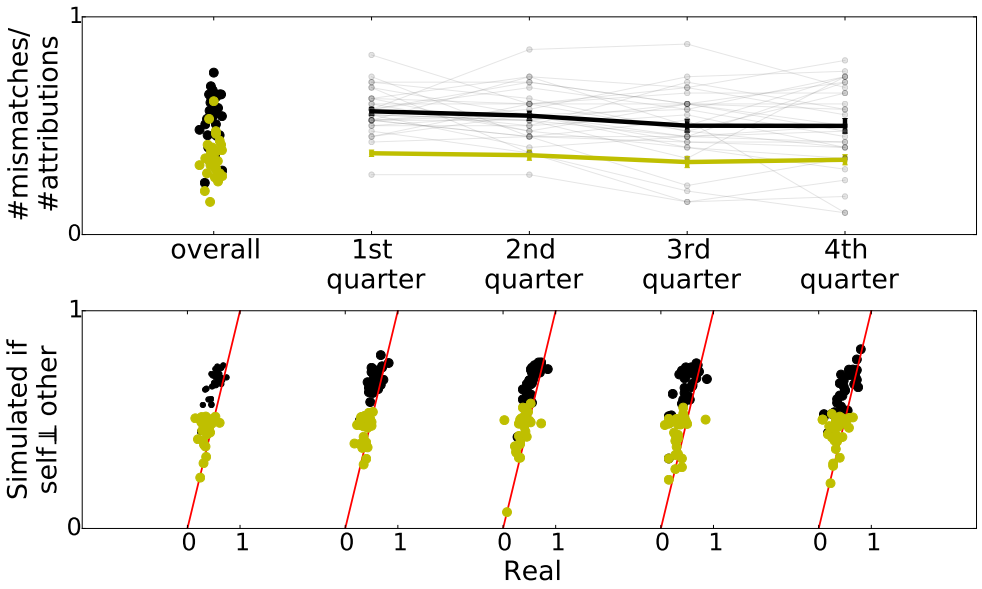


Figure 4.4: Differences in attribution responses between the two conditions. Black is used for differences computed using all 4 attribution options, yellow is used for differences computed after relabelling them as 'internal' vs 'external'. Top: real proportion of differences, overall (left) and across time (right). Dots and faded lines represent individual subjects. Think lines represent averages $\pm$ s.e.m across subjects. Bottom: the real proportion of differences vs the one expected if attributions in the two conditions were independent, each response being drawn from the subject's distribution of preferences for the 4 options; left: overall, right: across time.

4.2 Model agnostic analyses

In this section we present the results of our model agnostic analyses, aimed at identifying the effect that several factors of interest and their interactions have on attributions. These factors are outcome, objective task measures, objective performance measures and skill estimates.

Outcome was of interest to us because it is the most salient source of information about their performance to which subjects have access, but also because, in past studies, it has been repeatedly found to be related to subjects' causal attributions. Normal controls have been found to be biased toward making internal attributions for positive outcomes and external attributions for negative ones, as opposed to depressed patients, who displayed no such biases, in accordance with the "depressive realism" theory (Alloy

and Abramson, 1979; Martin et al., 1984; Vázquez, 1987; Bentall and Kaney, 2005; Campbell and Sedikides, 1999) (see also the literature review in chapter 1). We expected to find the same self-enhancement bias in our subjects.

Objective task measures and performance measures were factors of interest because they provided information about the aspects of the task referenced in the available attribution options, and any rational response strategy should reflect their influence. We do not propose a normative account of precisely how subjects should integrate these inputs into their causal attributions. Rather, our aim is to establish whether subjects' attribution responses reflect reasonable use of available information, in which case effects of objective task measures on attribution can be used as references against which to compare any effects of skill estimates. Finally, previous skill responses were the main

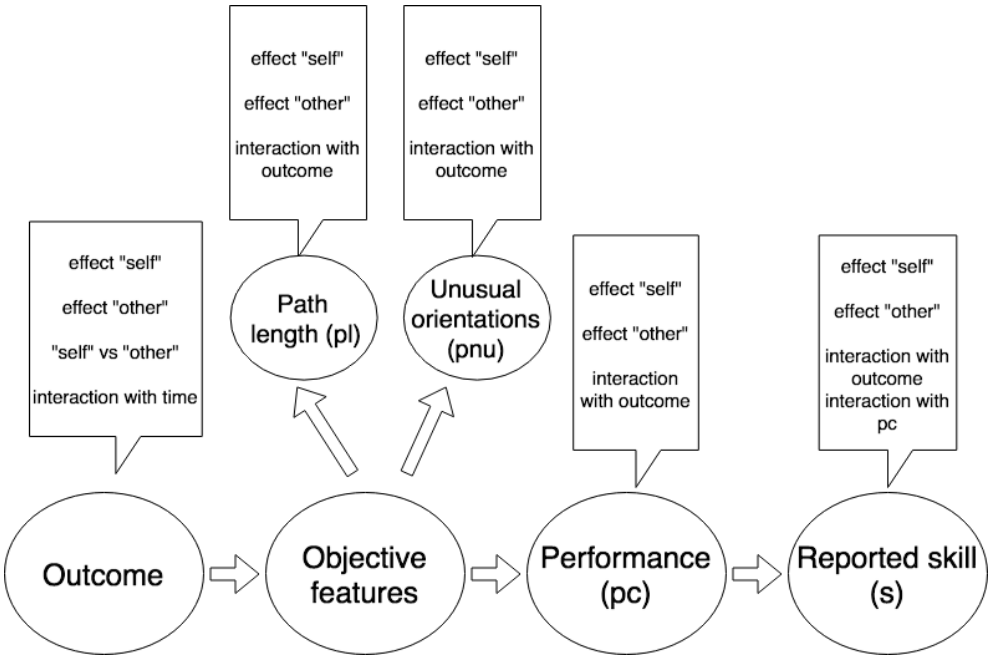


Figure 4.5: Roadmap of model-agnostic analyses.

factor of interest, in accordance with our overarching goal of investigating reciprocal influences between beliefs about the self and attributions.

We begin with a short presentation of the techniques we used for these model agnostic analyses, the challenges we encountered, and caveats to keep in mind, then proceed to presenting the results (see figure 4.5 for a visual

roadmap). All subsections - each dedicated to one of the factors of interest - share the same structure: we begin by briefly stating our expectations or hypotheses we set out to test about the factor, if any; we illustrate the effects we observed in the data; finally we report the results of statistical tests and discuss them briefly[1].

4.2.1 Technical aspects

Our dependent variables of interest were the attribution responses. For the purposes of the analyses presented in this section, based on statistical tests which involved no trial by trial modelling, we needed summary statistics of these data.

For a given level of a factor of interest and a given attribution option, the summary statistic we used was the proportion of attributions to the option, out of all attributions provided for the factor level: e.g. for the effect of outcome on internal attributions in the "self" condition we compared the proportions of attributions to self out of all attributions provided for wins, vs the proportion of attributions to self out of all attributions provided for losses.

Factors of interest other than outcome were continuous, and we chose to test for their effect by discretising. Due to the fact that different factors provide evidence for competing response options, we expected u-shaped effects of individual factors on attributions to corresponding options: values closer to the extremes of the range could be salient enough to significantly increase (or decrease) the likelihood of attributions to a given option, while intermediate values would produce smaller effects (if any), as the factor influence would in this case be diluted into the contribution of competing factors. In order to be able to detect such u-shaped or inverse u-shaped effects, we chose a quartile discretisation. For each subject and each factor of interest, discretisation was performed on the z-scored factor values.

To test for the effects of interest, we performed permutation tests using

[1] In these model agnostic analyses we investigated each factor separately; see appendix L for information about correlations between them.

either repeated measures t-test statistics, or F statistics associated with the relevant one or two-way repeated measures ANOVA (Howell, 2012) (see appendix O); reported p-values are estimates from these permutation tests. We used the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) to control FDR at level 0.05, and all results reported as significant survive the correction, unless otherwise stated.

As described above, a measurement of the attribution variable for a given subject and a given level of a factor of interest was the proportion of attributions for a given option, computed over the relevant trials. As a result, different measurements had different associated uncertainties, as if our measurements were not single measurements, performed with the same instrument, but averages over different number of raw readings. There are several reasons why the numbers of relevant trials, and therefore the associated uncertainties, can differ. For the case of outcome as the factor of interest, the staircase procedure was aimed at maintaining roughly equal numbers of wins and losses throughout the task, but it did not provide perfect equality between the number of wins and losses, nor could it guarantee that the number of validly expressed attributions would be equal for wins and losses; in practice, differences between these numbers did occur. The same is true for other factors, where quartile quantisation resulted in close, but sometimes not identical, numbers of trials per quartile. The issue is also present in analyses of the interaction between outcome and other factors: due to the coarseness of the staircase adaptation mechanisms (see chapter 2), correlations between levels of objective task measures and outcomes were not entirely removed, resulting in e.g. fewer wins than losses for high levels of the correct path length variable (see figures in appendix P).

Not accounting for these measurement uncertainties when simply performing t-tests or ANOVA might produce misleading results; however our choice of using permutation tests to determine statistical significance guards against this possibility. We did not account for uncertainties in the F-statistics

computations (see appendix O for detailed information about the repeated measures F-statistics that we used), which would have required more elaborate corrections, but we did account for them in computing t-statistics[2]. See appendix M for details.

4.2.2 Outcome

Previous studies (Alloy and Abramson, 1979; Tillman and Carver, 1980; Martin et al., 1984; Vázquez, 1987; Bentall and Kaney, 2005), (see Campbell and Sedikides, 1999; Mezulis et al., 2004, for reviews) repeatedly found that normal subjects have a propensity to make internal attributions for positive outcomes and external attributions for negative ones, despite having exercised the same level of control over both types of outcome; we expected to find the same effect in our task.

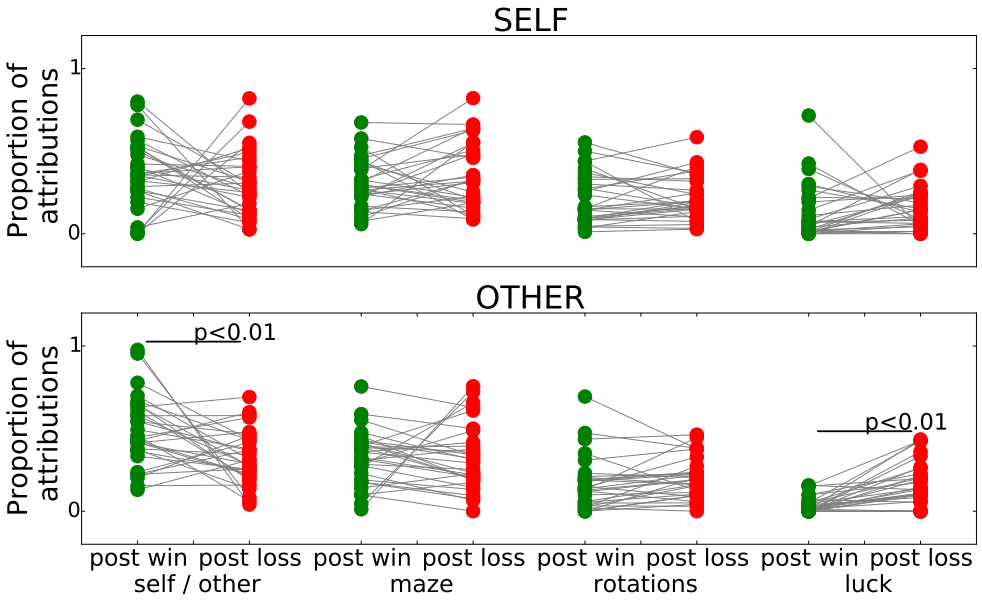


Figure 4.6: Effect of outcome on attributions to the 4 available options. Each line represents a subject, red is used for attributions post loss, green for attributions post win. P-values from paired t-tests, see text.

However, in our dataset, even though subjects tended to make more internal attributions for wins than for losses, the effect was not significant in the

[2]Note that such corrections are not necessary for model-dependent analyses (see section 4.3), which deal with the issue of varying number of trials for the different levels of the factors of interest by providing trial-by-trial predictions and likelihoods.

"self" condition (paired $t(15) = 0.7$, p-value from permutation test $p = 0.25$). However, we note that there were two subjects who made no attribution to self for wins[3], and that excluding these resulted in a significant effect of outcome on the proportion of attributions to self for wins vs losses (paired $t(13.5) = 2.38$, $p = 0.01$). In the "other" condition subjects also attributed more wins to the "other" than they attributed losses, and the effect was significant (paired $t(15) = 4.02$, $p < 0.01$). Direct comparisons between the two conditions showed that subjects made more internal attributions in the "other" condition for both wins and losses, with the difference being significant for wins (paired $t(15) = 3.62$, $p < 0.01$), but not for losses (paired $t(15) = 0.95$, $p = 0.35$).

Further tests for the effect of outcomes on specific external attributions to maze, rotations and luck showed no significant effect in the "self" condition[4], and significant increases in attributions to rotations (paired $t(15) = 1.95$, $p = 0.02$) and luck (paired $t(15) = 1.93$, $p < 0.01$) after losses compared to wins in the "other" condition (see figure 4.6; for the permutation distributions of the statistics see figures in appendix N).

Due to the nature of the task, which was framed in terms of learning, we expected to find effects of time, and interactions between time and outcome might be related to the weakness of the effect of outcome on internal attributions in the "self" condition.

Figures 4.7 and 4.8 suggest that interactions between time and outcome were indeed present. We note that, surprisingly, subjects made more internal attributions for negative outcomes than for positive ones at the beginning, and that this preference switched during the task, leading to the opposite preference by the end of the task. This was valid for both conditions, which means that at the beginning of the "other" condition, which always followed

[3]We report results of tests excluding these two subjects when they are different from the results of tests performed on data from all subjects.

[4]However excluding the two subjects who made no attribution to "self" for wins, we found a significant increase in attributions to luck after losses compared to wins (paired $t(13.5) = 1.62$, $p = 0.01$).

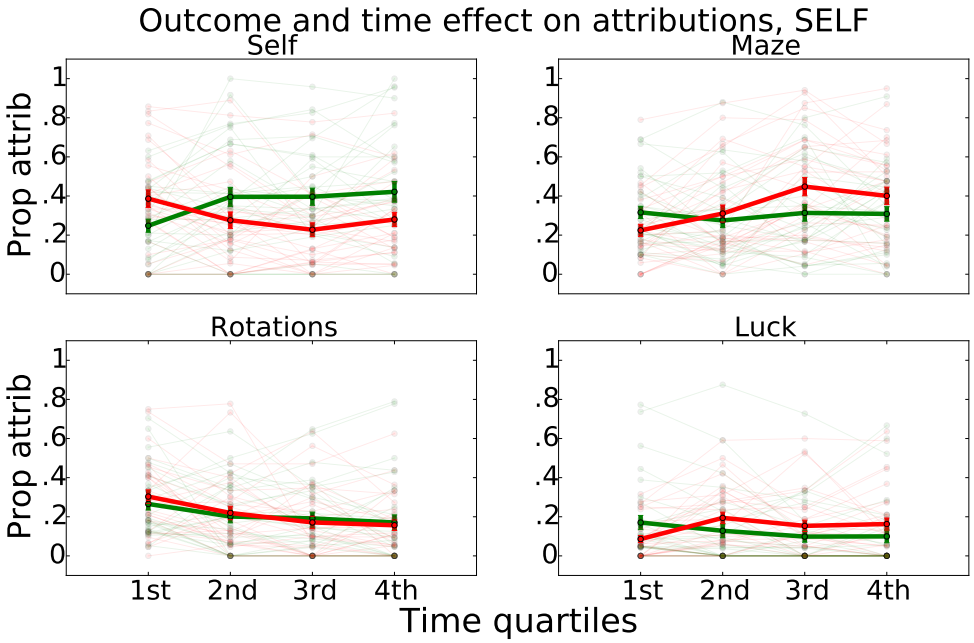


Figure 4.7: Effect of outcome and time on attributions in the "self" condition. Faded lines: individual subjects; thick lines: mean ± s.e.m across subjects. Greeen: attributions for wins, red: attributions for losses.

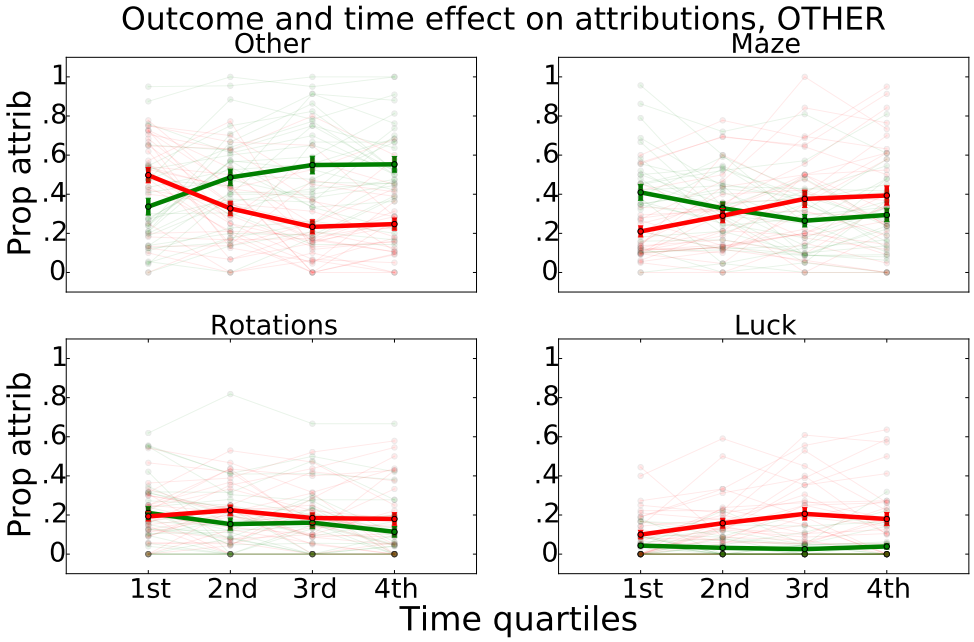


Figure 4.8: Effect of outcome and time on attributions in the "other" condition. Faded lines: individual subjects; thick lines: mean ± s.e.m across subjects. Greeen: attributions for wins, red: attributions for losses.

the "self" condition, subjects' responses again showed a tendency to attribute negative outcomes internally, and the preference again flipped by the end of the trials. This would suggest that, contrary to what we might expect from previous research, people do not start from a default preference of making internal attributions for positive outcomes.

As far as attributions to "rotations" are concerned, they seemed to be decreasing, irrespective of outcome; this would be consistent with subjects perceiving (or expecting) rotations to have less impact on their performance as their skill improves.

We tested these effects by permutation tests, using the F-statistics employed in repeated measures two-way ANOVA, with outcome and time as fixed factors and subjects as random factors; see appendix O for a detailed description of the permutation tests. In the "self" condition we found significant interactions between outcome and time for internal attributions ($F_{(3,90)} = 13.46, p < 0.01$) and for attributions to "maze" ($F_{(3,90)} = 6.98, p < 0.01$) and "luck" ($F_{(3,90)} = 7.35, p < 0.01$), and no significant interaction for attributions to "rotations" ($F_{(3,90)} = 0.81, p = 0.49$), but a significant effect of time for attributions to "rotations" ($F_{(3,90)} = 7.22, p = 0.01$). In the "other" condition, we also found significant interaction effects for internal attributions ($F_{(3,90)} = 28.72, p < 0.01$) and attributions to "maze" ($F_{(3,90)} = 15.25, p < 0.01$) and "luck" ($F_{(3,90)} = 5.41, p < 0.01$), along with a significant main effect of outcome for internal attributions ($F_{(1,30)} = 9.21, p < 0.01$), and a main effect of time for attribution to rotations which did not survive multiple comparison corrections ($F_{(3,90)} = 3.54, p = 0.0145$).

For a direct comparison between the two conditions, see figure 4.9, which shows average attribution proportions for "self" and "other" in the same figure.

As mentioned above, subjects made more internal attributions in the "other" than in the "self" condition; this difference declined for losses, but increased for wins (post hoc tests, internal attributions for wins "self" vs

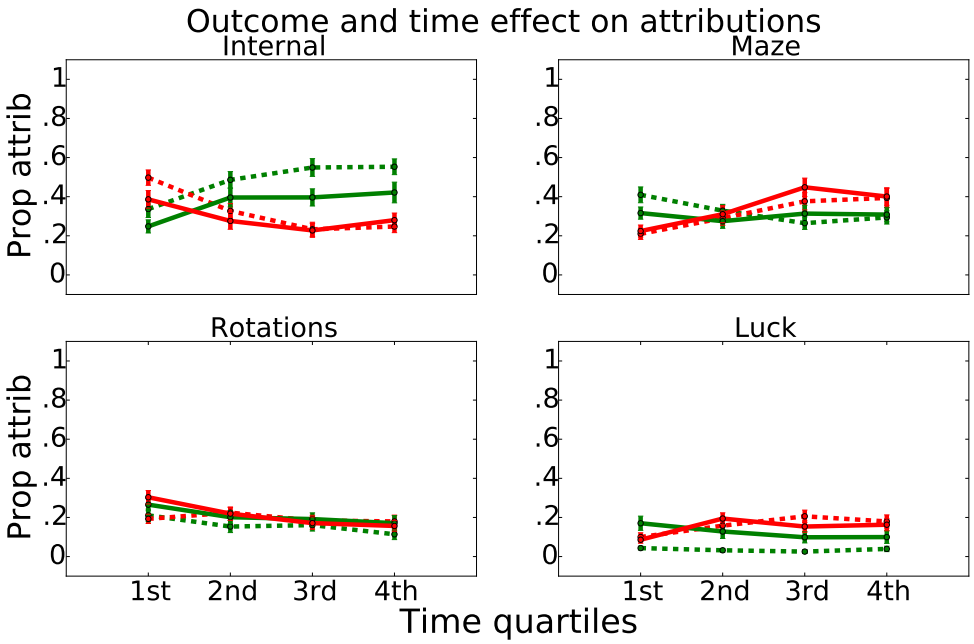


Figure 4.9: Effect of outcome and time on attributions, comparison between the "self" and "other" conditions. Mean ± s.e.m across subjects. Green: wins, red: losses. Plain line: self, dotted line: other.

"other": 1st quarter paired $t(15) = -1.92, p = 0.06$, 2nd quarter paired $t(15) = -1.91, p = 0.06$, 3rd quarter paired $t(15) = -3.5, p < 0.01$, 4th quarter paired $t(15) = -2.9, p < 0.01$). Along with higher average skill estimates for "other" (see 3), this pattern is consistent with people being "nicer" to others than to themselves, which has been observed in other contexts (Crockett et al., 2014; Rand et al., 2014) (see Rand and Nowak, 2013, for a review). However, the fact that there are also, initially, more internal attributions for losses in the "other" condition (post hoc test, internal attributions for losses "self" vs "other": 1st quarter paired $t(15) = -2.23, p = 0.04$) is not consistent with a general tendency of judging the "other" more favourably. We return to actor-observer effects and their relationship with outcome valence and self-serving biases in more detail in the discussion section at the end of this chapter (4.6).

4.2.3 Objective task and performance measures

We offered subjects different response options for external attributions in order to be able to test whether their responses reflected objective task measures. We do not propose a normative model of how this information should be integrated in subjects' attribution responses. Rather, our point is to establish that subjects' attributions reasonably reflected measurable task or performance variables; we present the effect of these measurable factors as a background and context for the effect of subjects' own skill estimates on attribution responses.

In this section we present the results of these analyses with regard to two objective task measures - the length of the correct path through the maze and the proportion of unusual orientations - and one performance measure - the average proportion of correct key presses.

4.2.3.1 Path length

We expected the length of the correct path through the maze to be linked to subjects' attributions to the "maze complexity" option. Specifically we expected that trials for which the path length was in the extremes of the path length distribution would be associated with more attributions for this option than trials with intermediate values of path length. This is indeed what we observed, for both conditions, see figure 4.10.

Rational attribution responses would also predict that,as the length of the correct path increases, subjects would make more attributions to maze for losses and less attributions to maze for wins, with the opposite pattern for internal attributions. Figures 4.11 and 4.12 show these expectations were confirmed, for both conditions.

We tested the significance of these effects by performing permutation tests, using ANOVA F-statistics. Because the bottom quantile of path length distribution was strongly associated with wins, and the top one strongly associated with losses (see figure P.1 in appendix P), we performed one-way ANOVA tests for path length only and restricted two-way ANOVA tests for

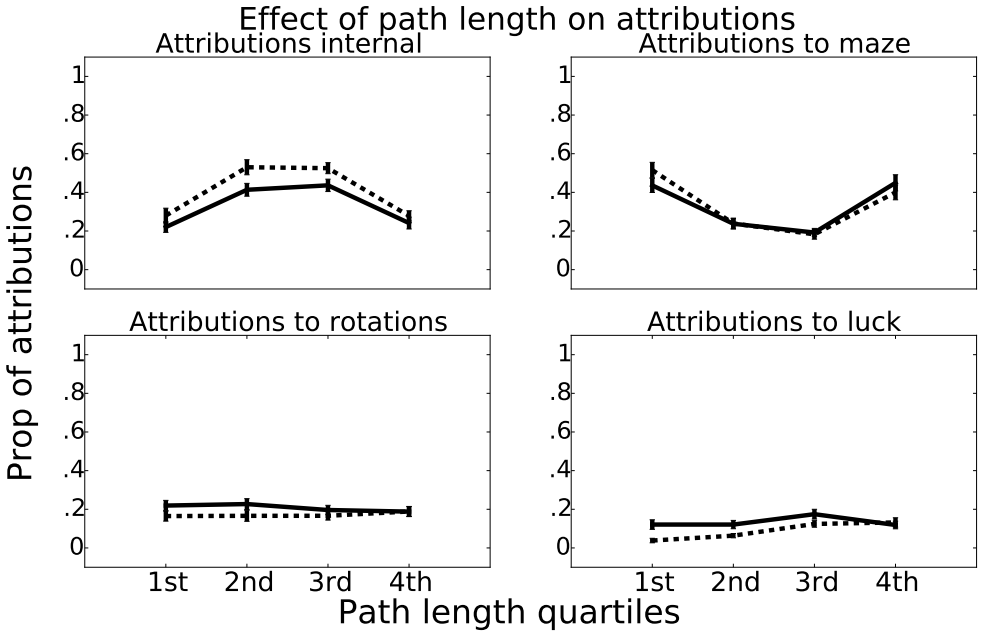


Figure 4.10: Effect of correct path length on attributions. Mean ± s.e.m across subjects. Plain line:"self", dotted line: "other".

path length and outcome to the intermediate quantiles of the path length distributions.

Permutation tests using the F-statistic from one-way repeated measures ANOVA with path length as a fixed factor confirmed that the effects of path length on internal attributions ("self": $F_{(3,90)} = 16.68, p < 0.01$, "other": $F_{(3,90)} = 26.59, p < 0.01$) and attributions to maze ("self": $F_{(3,90)} = 25.6, p < 0.01$, "other": $F_{(3,90)} = 33.46, p < 0.01$) were indeed significant. We also tested for effects on the other attribution options and found no significant effect on attribution to rotations in either condition, but an effect on attributions to luck, which was significant in the "other" condition ($F_{(3,90)} = 11.26, p < 0.01$), but did not survive multiple comparisons correction for the "self" condition.

Our hypothesis that the effects on internal attributions and attributions to maze were due to an interaction between outcome and path length were also confirmed by permutation tests using two-way repeated measures ANOVA with outcome and path length as fixed factors, performed only for the two intermediate path length levels. We found significant interactions between

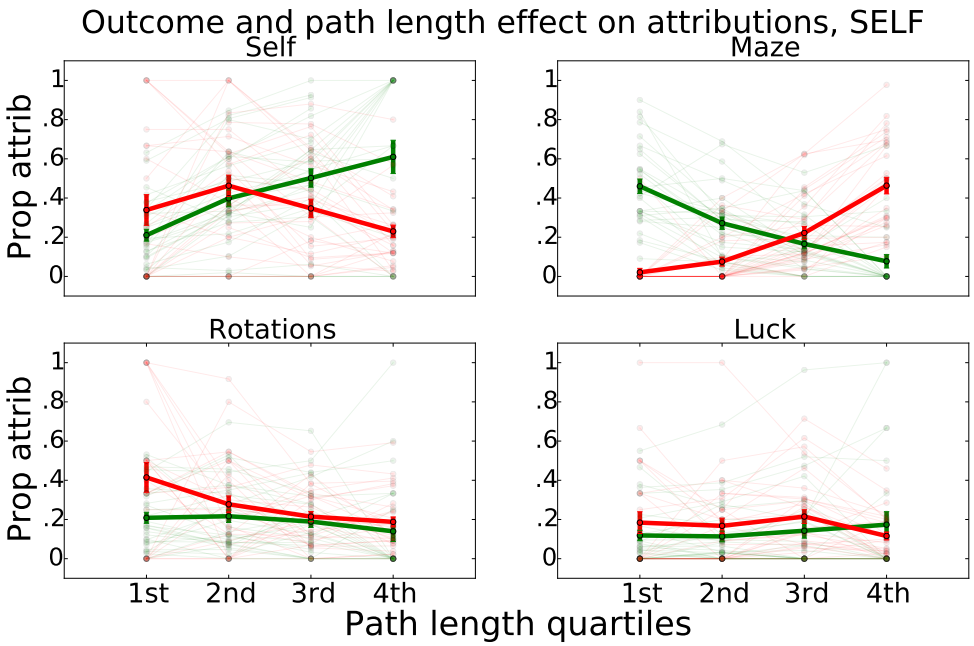


Figure 4.11: Effect of correct path length and outcomes on attributions, "self" condition. Faded lines: individual subjects; thick lines: means ± s.e.m across subjects. Green: wins, red: losses.

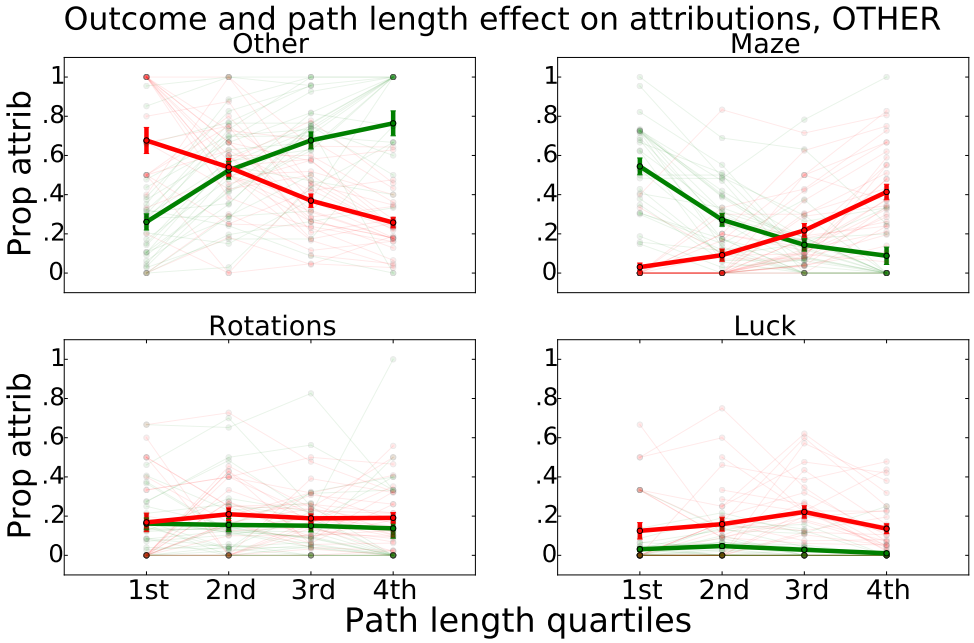


Figure 4.12: Effect of correct path length and outcomes on attributions, "other" condition. Faded lines: individual subjects; thick lines: means ± s.e.m across subjects. Green: wins, red: losses..

outcome and path length on internal attributions ("self": $F_{(1,30)} = 17.85, p < 0.01$, "other": $F_{(1,30)} = 30.67, p < 0.01$) and on attributions to "maze" ("self": $F_{(1,30)} = 45.14, p < 0.01$, "other": $F_{(1,30)} = 35.83, p < 0.01$) in both conditions, and no surviving main effect. As far as the other response options were concerned, no effect or interaction survived multiple comparisons corrections for attributions to rotations in either condition; the same is true for attributions to luck in the "self" condition. For the "other" condition, we found a significant effect of outcome on attributions to luck ($F_{(1,30)} = 27.1, p < 0.01$), which, together with the skewed distribution of outcomes per path length quantiles, explains the effect of path length on attributions to luck in the "other" condition, mentioned above.

These results prove that subjects' attribution responses were sensitive to the path length manipulations, and suggest that subjects rationally integrated path length and outcome information in their attribution responses.

4.2.3.2 Proportion of non-up orientations

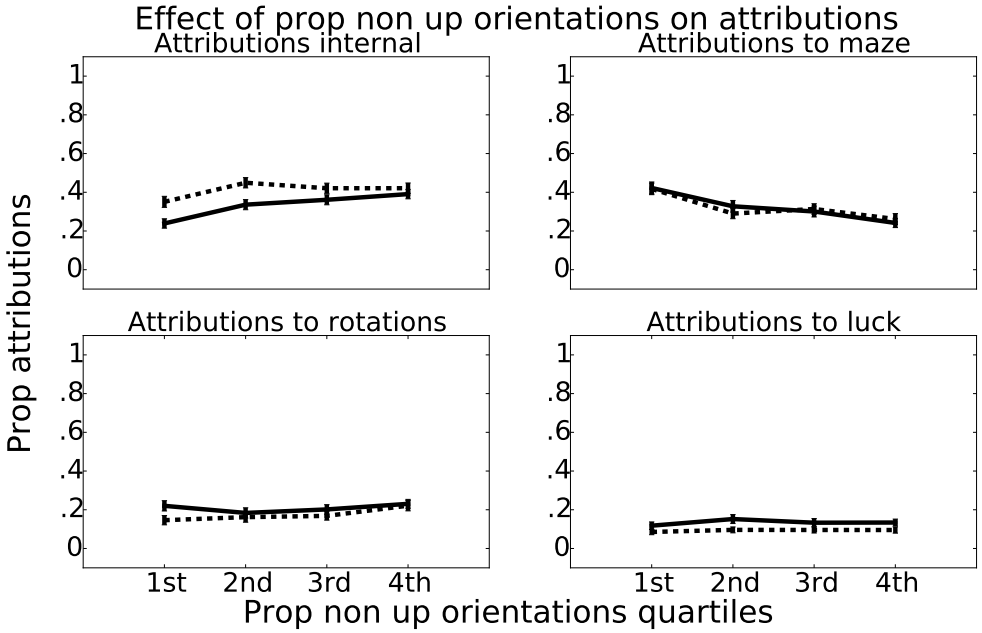


Figure 4.13: Effect of non-up maze orientations on attributions. Means ± s.e.m across subjects. Plain line: "self", dotted line: "other".

As described previously in 2.3.2, we used the proportion of unusual (non-

up) orientations of the maze within a trial (pnu) as a measure of the contribution of rotations to the difficulty of the trial. We expected this variable to be linked to subjects' attributions to rotations in a pattern similar to the one we postulated to connect path length with attributions to maze and internal attributions. Specifically, we expected higher proportions of attributions to rotations for the extreme quantiles of the prop non up orientations, and lower attribution proportions for intermediate levels of this variable. We also expected the opposite pattern for internal attributions.

Unlike the case of the path length variable, however, this is not what we observed (see figure 4.13). Subjects made fewer internal attributions for trials in the bottom quartile of the pnu variable, but this was not the case for the top quartile. There was also no u-shaped pattern for attributions to rotations. There was, instead, a decrease in attributions to maze for increasing proportions of unusual orientations, in both conditions.

We also expected interactions with outcome, with more unusual orientations leading to more internal attributions for wins and more attributions to rotations for losses. Figures 4.14, 4.15 show the effect of outcome and proportion of non up orientations on subjects' attribution responses in the two conditions.

We performed permutation tests on the F statistic from a 2-way repeated measures ANOVA with outcome and pnu as fixed factors. These tests indicated that interactions between outcome and pnu were significant in the "self" condition, both for internal attributions ($F_{(3,90)} = 6.76, p < 0.01$) and for attributions to rotations ($F_{(3,90)} = 17.9, p < 0.01$). In the "other" condition, the interaction was significant for attributions to rotations ($F_{(3,90)} = 4.65, p < 0.01$), but not for internal attributions. In addition, for both condition there was a significant main effect of pnu on internal attributions ("self": $F_{(3,90)} = 19.09, p < 0.01$, "other": $F_{(3,90)} = 19.7, p < 0.01$) [5].

[5] In addition, we found other main effects and interactions on attributions to maze and luck, in both conditions, which we had not made any prediction about, as we do not have a normative model of how subjects should distribute their attributions among the various external options, in cases where none of the options is particularly salient.

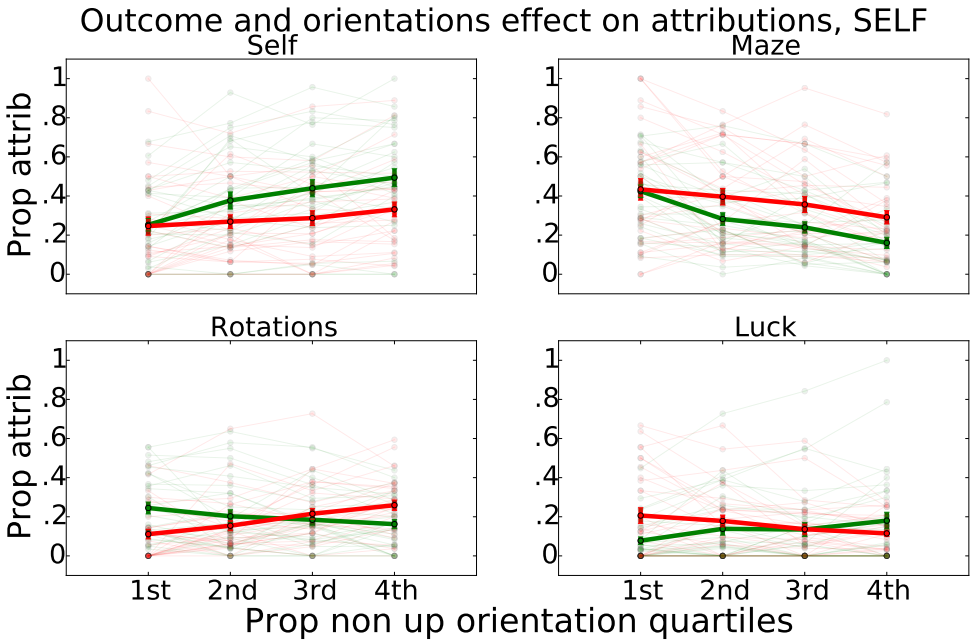


Figure 4.14: Effect of non-up maze orientations and outcomes on attributions, "self" condition. Faded lines: individual subjects; thick lines: means ± s.e.m across subjects. Green: wins, red: losses.

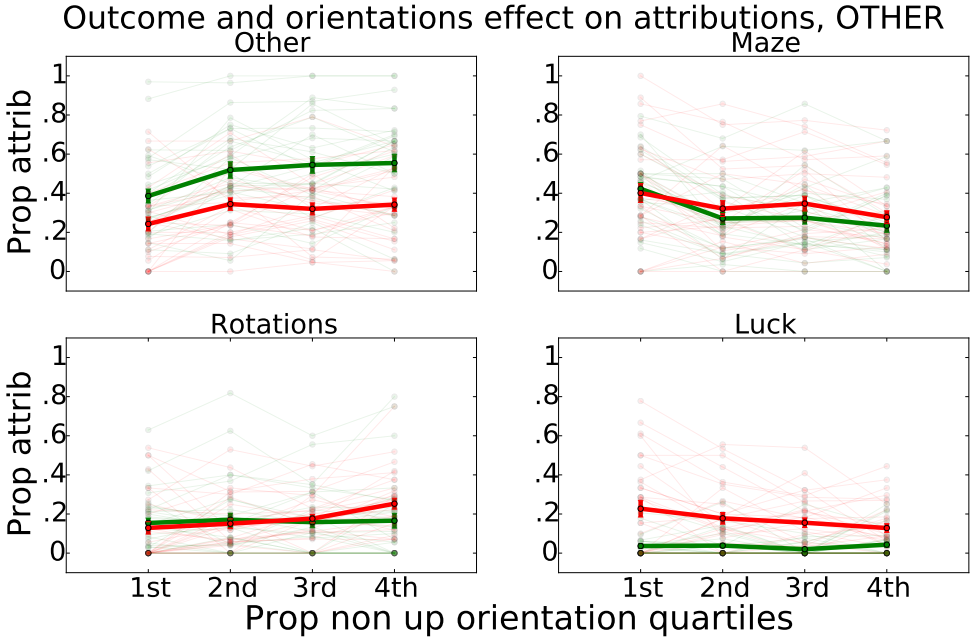


Figure 4.15: Effect of non-up maze orientations and outcomes on attributions, "other" condition. Faded lines: individual subjects; thick lines: means ± s.e.m across subjects. Green: wins, red: losses

These results were only partially in accord with our predictions. As we expected, subjects took more credit for wins with increased pnu; however we also expected they would more strongly avoid responsibility for losses, blaming them instead on rotations, which was not the case. Contrary to our expectations, the patterns of effects on internal attributions were different between path length and orientation. In addition, orientation information seems to have been processed differently for "self" and "other".

4.2.3.3 Objective performance

We tested for subjects' sensitivity to their own objective performance in their attribution responses by considering the effects of the proportion of correct key presses (pc) on attribution responses.

Subjects should make more internal attributions for wins and fewer internal attributions for losses with increased performance, if they are able to monitor their performance and if they use this information rationally. It is less clear what a signature of rational performance evaluation would involve in terms of attributions to the other available options: we predicted that performance would be reflected in subjects' attributions to luck, in that the better their performance, the more luck is to be blamed for losses, and the worse their performance, the more luck is responsible for wins. However subjects could, alternatively, assign losses associated with good performance to the task difficulty, either to rotations or to maze, according to whichever aspect of the trial was more salient.

Figures 4.16 and 4.17 show the average attribution proportions as a function of outcome and pc in the two conditions. In both conditions, higher performance was associated with decreased internal attributions for losses, increased internal attributions for wins and increased attributions to luck for losses. For "self", attributions to luck for wins decreased with performance, which was not the case for "other", perhaps due to a floor effect.

We performed permutation tests using the 2-way repeated measures ANOVA F-statistic and found a significant interaction between outcome

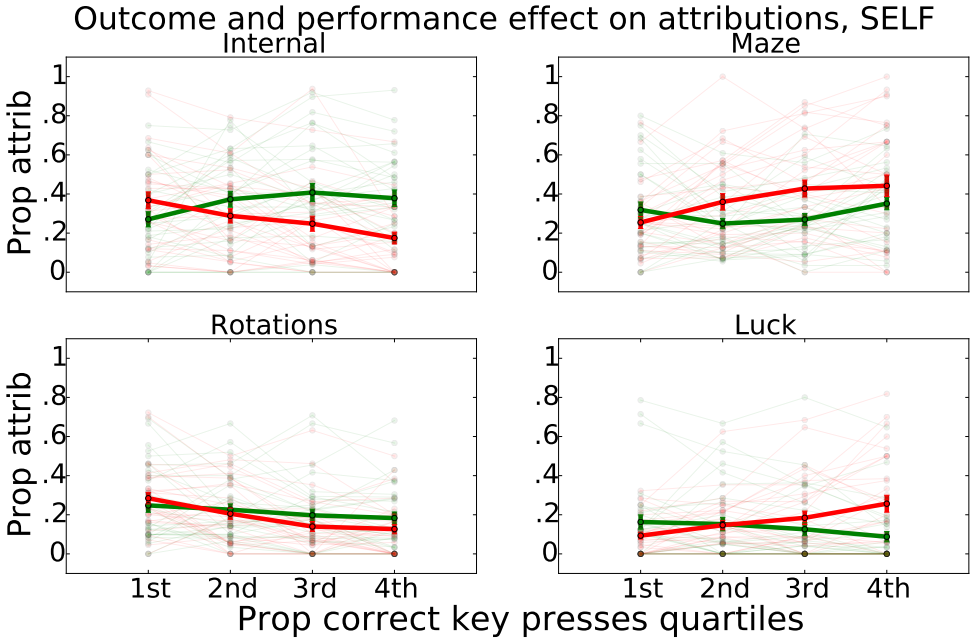


Figure 4.16: Effect of outcomes and proportion of correct key presses on attributions, "self" condition. Faded lines: individual subjects; thick lines: means ± s.e.m across subjects.

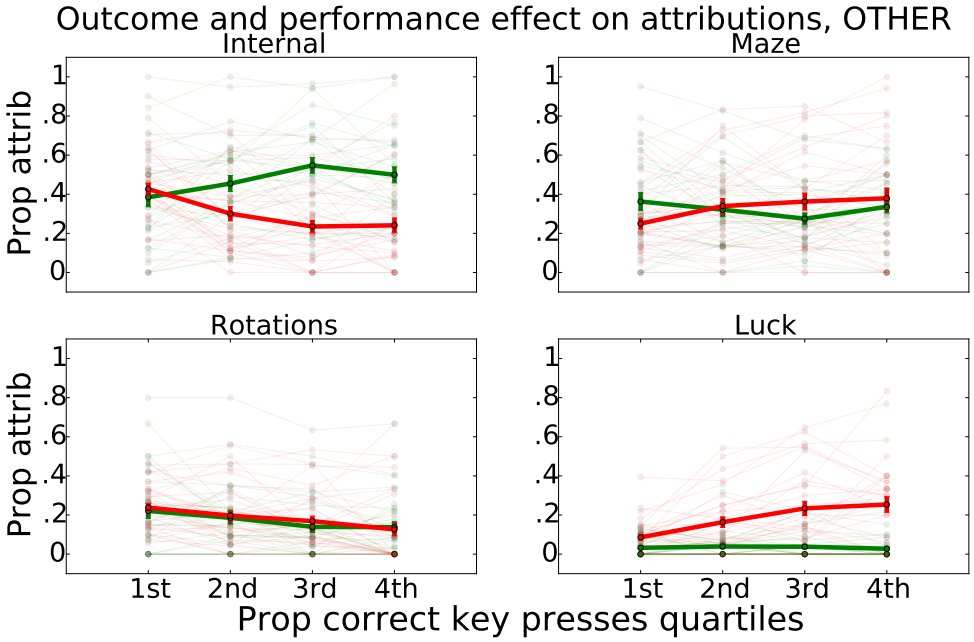


Figure 4.17: Effect of outcomes and proportion of correct key presses on attributions, "other" condition. Faded lines: individual subjects; thick lines: means ± s.e.m across subjects.

and pc for internal attributions ($F_{(3,90)} = 13.73, p < 0.01$) and attributions to luck ($F_{(3,90)} = 13.17, p < 0.01$) in the "self" condition, and no significant main effects of the outcome or performance measure on these attributions. In the "other" condition there was a main effect of outcome ($F_{(1,30)} = 11.36, p < 0.01$), as well as a significant interaction between outcome and pc ($F_{(3,90)} = 26.49, p < 0.01$), on internal attributions. As far as attributions to luck are concerned, both main effects (outcome: $F_{(1,30)} = 35.43, p < 0.01$, pc: $F_{(3,90)} = 11.89, p < 0.01$) and the interaction ($F_{(3,90)} = 12, p < 0.01$) were significant, due to subjects making almost no attributions to luck for wins in this condition, and making increasing numbers of attributions to luck for losses with increasing performance.

In addition, in both conditions attributions to rotations significantly decreased with increased performance, irrespective of outcome, ("self": $F_{(3,90)} = 10.78, p < 0.01$, "other": $F_{(3,90)} = 13.33, p < 0.01$), consistent with the fact the correct key presses performance measure is related to subjects' ability to deal with rotations.

These results showed that, both when evaluating themselves and when evaluating the "other", subjects were sensitive to performance, and integrated this information along with the outcome in their reports of causal attributions.

4.2.4 Previous skill estimates

Our main question of interest concerning analyses of attribution responses was wether subjects' beliefs about their own skill had an effect on the way they made causal attributions.

Other than being instructed to respond truthfully, subjects were in no way incentivised to tell the truth in their attribution responses. However, model agnostic analyses reported above showed that subjects' attribution responses displayed reasonable effects of task manipulations and their own performance, giving us reasons to believe that subjects' responses were meaningful and reflected plausible internal beliefs. We therefore assume that the influences of their beliefs about their skill (as reflected in their skill estimates) on their

attributions can also be revealed be the same approach.

Contrary to what we expected based on previous research (Alloy and Abramson, 1979; Martin et al., 1984; Vázquez, 1987; Bentall and Kaney, 2005) (see Campbell and Sedikides, 1999, for a review), subjects did not display a general pattern of making more internal attributions for positive outcomes than for negative ones (see figures 4.18 and 4.19) . Instead, their attributions were consistent with their skill responses, with negative internal attributions for low skill levels, and positive ones for higher skill levels. The opposite pattern was present for attributions to maze. However the switch between more internal attribution for losses and more internal attributions for wins (and the opposite for attributions to maze) happened at low levels of skill estimates, suggesting that subjects do indeed display self-serving biases.

Permutation tests with outcome and skill estimate level as fixed factors in a two-way repeated measures ANOVA revealed significant interactions between outcome and skill for internal attributions ($F_{(3,90)} = 9.08, p < 0.01$) and attributions to maze ($F_{(3,90)} = 6.65, p < 0.01$) in the "self" condition. Attributions to luck also showed a switch similar to the one for attribution to mazes, and attributions to rotations decrease with increasing skill estimates, but these two effects were not significant in the "self" condition, according to our permutation tests.

Surprisingly, but consistent with all previous analyses of skill estimates, these effects of skill appear to be stronger in the "other" condition. In this case as well subjects made fewer internal attributions for losses and more internal attributions for wins as skill estimates increased, with an early cross between the two curves. The opposite patterns were present in attributions to maze and rotations, with an early cross between the descending curve for maze attributions for wins and the ascending curve for maze attributions for losses, and no cross for the rotation attribution curves. Attributions to luck for losses increased with skill estimates, while attributions to luck for wins were at floor irrespective of the skill level.

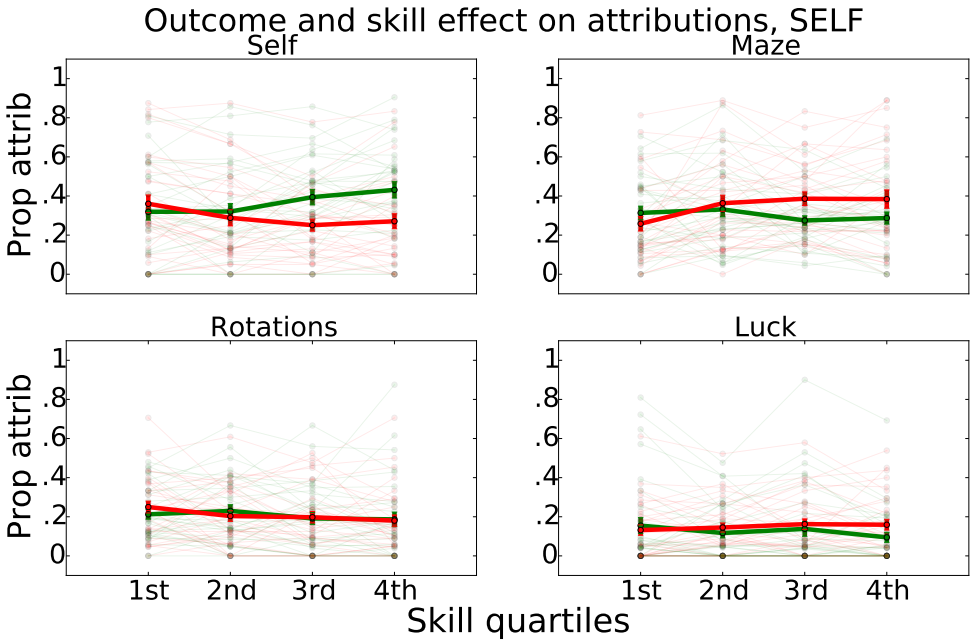


Figure 4.18: Effect of outcomes and skill responses on attributions, "self" condition. Faded lines: individual subjects; thick lines: means ± s.e.m across subjects. Green: wins, red: losses.

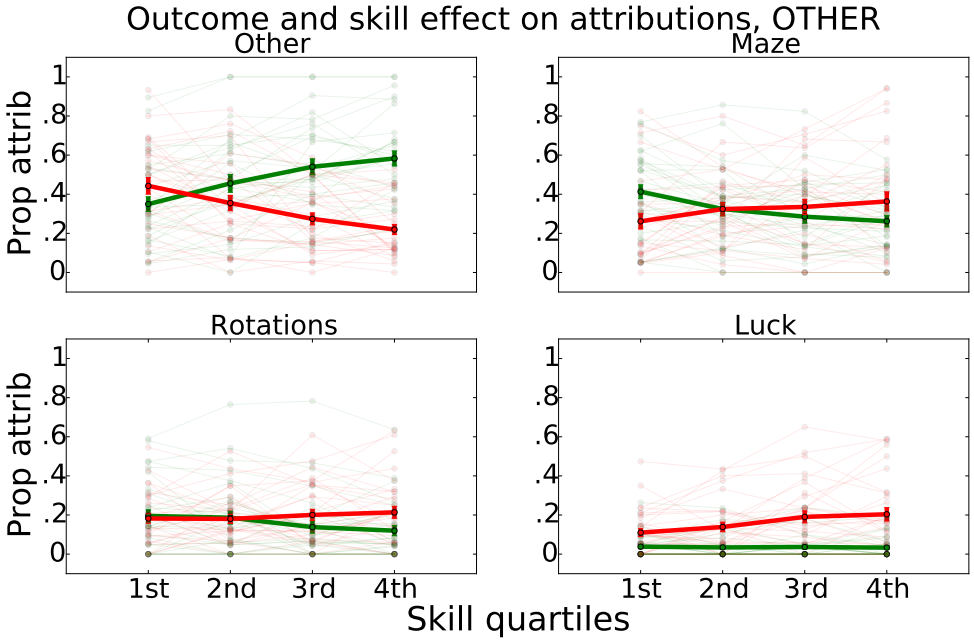


Figure 4.19: Effect of outcomes and skill responses on attributions, "other" condition. Faded lines: individual subjects; thick lines: means ± s.e.m across subjects. Green: wins, red: losses.

Permutation tests for this condition identified significant interactions between outcome and skill estimates on all attributions ("other": $F_{(3,90)} = 38.84$, $p < 0.01$, "maze": $F_{(3,90)} = 10.39, p < 0.01$, "rotations": $F_{(3,90)} = 5.49, p < 0.01$, "luck": $F_{(3,90)} = 5.7, p < 0.01$), but also significant main effects of outcome on internal attributions ($F_{(1,30)} = 9.65, p < 0.01$), and on attributions to luck ($F_{(1,30)} = 29.54, p < 0.01$).

In order to check whether subjects' beliefs about skill had any effect on internal attributions over and above objectively measured performance, we also performed permutations tests for wins and losses separately, using the statistics corresponding to repeated measures two-way ANOVAs with skill and performance as fixed factors. Due to the small number of trials, for the purposes of this analysis we quantised skill and performance in tertiles rather than quartiles.

Figures 4.20 and 4.21 show the proportion of internal attributions as a function of the performance level for each level of reported skill, separately for wins and losses, for "self" and "other" respectively. These plots suggest that skill estimates have an effect on internal attributions, over and above that due to performance. Permutation tests performed to check the significance of these effects revealed an effect of skill in the "self" condition for losses which did not survive multiple comparisons corrections, and significant effects of skill for both wins ($F_{(2,60)} = 20.56, p < 0.01$) and losses ($F_{(2,60)} = 7.43, p < 0.01$) in the "other" condition.

Our question of interest refers to relationships between self-beliefs and causal attributions, so our purpose in this analysis was to investigate whether the effect of reported skills on internal attributions could be reduced to the effect of performance. We found that at least in the "other" condition this is not the case, and that beliefs about skill exerted an influence on internal attributions over and above performance. While the effect was not significant in the data from the "self" condition, this might be due to the higher level of noise in this condition, rather than the absence of the effect. These results do not ex-

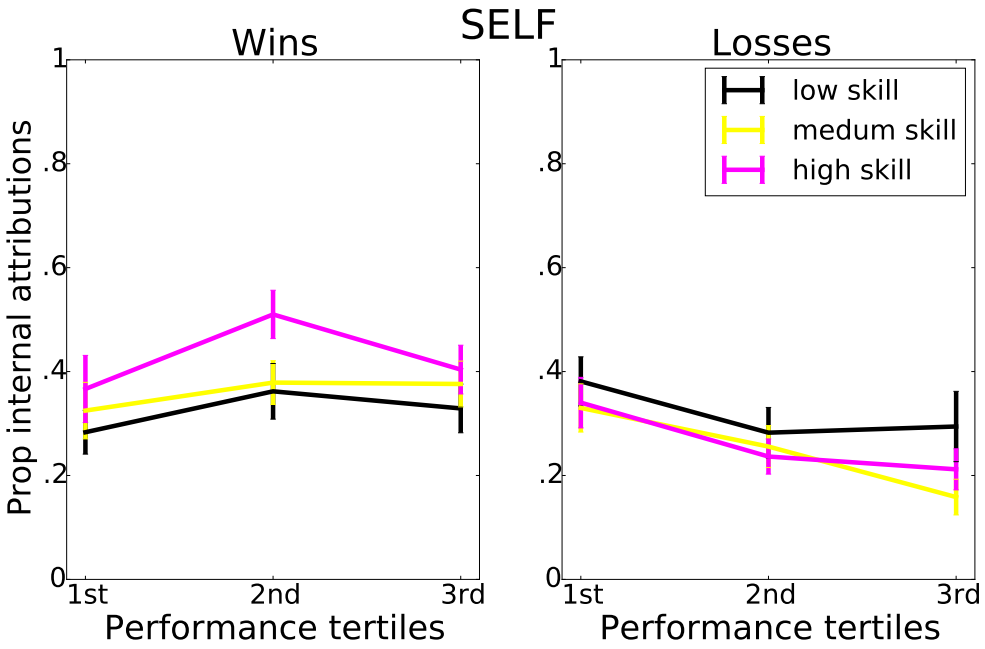


Figure 4.20: Effect of skill responses and performance on attributions, "self" condition. Means ± s.e.m across subjects.

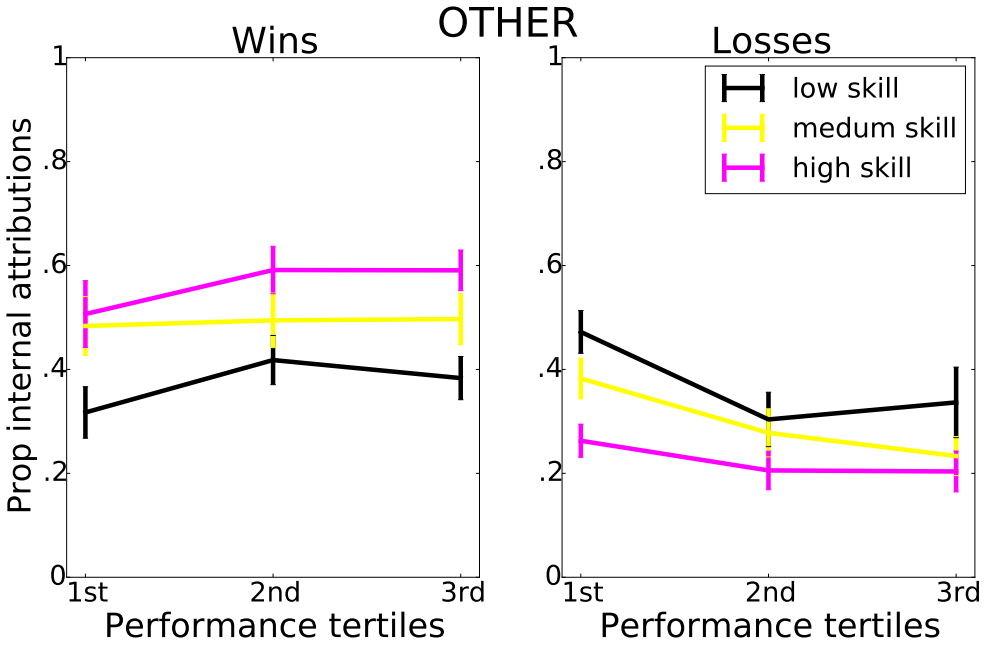


Figure 4.21: Effect of skill responses and performance on attributions, "self" condition. Means ± s.e.m across subjects.

clude the possibility of a richer measure of performance being able to account for the effect of skill, however they do point to the fact that subjects' reported beliefs contained additional information about internal attributions, compared to simple objectively measurable aspects of performance. We provide more evidence for the skill-performance distinction in the later model-dependent analyses section (4.3.3.3).

4.2.5 Model agnostic analyses: summary

Model agnostic analyses presented in this section identified effects of outcome, task features, performance and skill estimates on attributions. Some of the identified effects confirmed out predictions, while others were unexpected, suggesting questions for future work.

As far as **outcome** is concerned, as predicted, subjects made more internal attributions for wins than for losses. This effect was significant for the "other" condition, as well as for the "self" condition, when excluding two subjects who made no internal attributions for wins. Direct comparison between "self" and "other" showed that subjects generally made more internal attributions in the "other" condition than in the "self" condition, with the effect being significant for wins, but not losses. These results are consistent with an actor-observer effect (Jones and Nisbett, 1987). They also suggest that in our data the expected self-serving bias is weaker than a bias which is "other-enhancing", indicating that subjects tend to be nicer to others than they are to themselves[6]. See section 4.6 for further discussion. Outcome

Interactions between outcome and time were significant for both conditions, however contrary to our expectations subjects started by making more internal attributions for losses, switching to the opposite pattern after some experience with the task. This pattern might be due to differences between our experimental context and the one of previous studies: our task was clearly framed in a learning context, which might have lead subjects to assume that

[6]It is important to note that in our case the pretended "other" is in fact the subject, and that it is difficult to establish to what extent and at what point subjects suspected this deceit.

they would start by being bad at the task, and to expect they would get better in time. We also note that in our experiment time and skill (both real and as estimated by subjects) are inherently to a large degree confounded, such that the pattern of preference changes in time might be due to subjects' changing beliefs about how good they are (see below).

Task measures

As far as the **objective task measures** are concerned, we expected path length and the proportion of unusual orientations (pnu) to have similar effects on internal attributions, and equivalent effects on attributions to maze and rotations respectively. Specifically, we expected that with increased path length (orientations) subjects would take more credit for wins and would accept less blame for losses, blaming them instead on maze (rotations).

Our expectations were confirmed for path length, but only partially for pnu: subjects did indeed take more credit for wins as pnu increased, but there was no decrease in internal attributions for losses associated with higher proportions of unusual orientations. In addition, the effects of pnu on attributions to rotations were qualitatively different for "self" and "other". This might reflect differences in the way they were perceived or processed. We conceived of path length and rotations as orthogonal potential causes, but subjects experienced rotations as nested within a frame provided by the maze, and might therefore perceive these factors as being hierarchically organised.

In addition, subjects learn to deal with rotations, and therefore their saliency is expected to evolve during the task. This is also true for path length, but to a lesser extent: due to our choices for the range of path lengths and available time and speed, there was very little subjects could do to improve their speed through the maze over and above reducing rotation-induced errors and or pauses. This difference can be interpreted in terms of different degrees of control over the two aspects of the task, which is particularly relevant for attribution.

Objective performance

As far as **objective performance** is concerned, we found the expected interactions with outcome regarding internal attributions and attributions to

luck. Increased performance was associated with subjects taking more credit for wins, accepting less responsibility for losses, and blaming bad luck more for losses, in both conditions.

We were particularly interested in relationships between **skill estimates** and attributions, and we model agnostic analyses showed that subjects' skill estimates had significant effects on their attributions for subsequent trials. We found the expected interactions between skill estimates and outcomes in attributions for both conditions: subjects took more credit for wins, and less responsibility for losses with increasing skill estimates, showing the opposite pattern in attributions to maze. Further analyses performed with skill and performance as factors indicated that skill influenced internal attributions over and above performance, the effect being significant for "other", although it did not survive multiple comparison corrections for "self". Skill estimates

4.3 Model-dependent analyses

In this section we present model-dependent analyses of attribution responses. The goal of these analyses was two-fold: to identify and compare the contributions of the different factors based on a more fine-grained, individual level trial by trial prediction, and to provide subject-level parameter estimates which can be analysed in connection with subjects' questionnaire responses. As before, we were particularly interested in the contribution of skill estimates in predicting attribution responses.

This section is structured as follows. We begin with a technical subsection presenting the models we compared, the fitting procedure, and the criterion used for model comparison. We then present the results of model comparisons. Finally, we present analyses of the best model parameters, focusing, as in the model-agnostic section, on the effects of the factors of interest: outcome, task and performance measures, skill estimates. The section ends with a summary.

4.3.1 Technical aspects

We used linear classification models. These assume that for each response option, a linear combination of the features of interest provides a score for that option; scores for the different options are then passed through a softmax function, to provide response probabilities for the different response options:

$$
\begin{aligned}
s_{t,o} &= \boldsymbol{w}_o \cdot \boldsymbol{f}_t \forall o \in O \\
p_t(o) &= \frac{\exp(s_{t,o})}{\Sigma_{o \in O} \exp(s_{t,o})}, \text{ where} \\
s_{t,o} &= \text{score of response option o on trial t} \\
\boldsymbol{w}_o &= \text{feature weights for option o} \\
\boldsymbol{f}_t &= \text{feature values on trial t} \\
O &= \text{set of available response options} \\
p_t(o) &= \text{probability of choosing option o on trial t}
\end{aligned}
\tag{4.1}
$$

The models we compared varied along two dimensions: the features included, and whether subjects were fitted independently or hierarchically. In terms of the features that were included, models belonged to one of the following six categories: no features - bias only; models including bias and one of the following set of features: reported skill, performance features, performance features and task features, reported skill and task features; and the full model: reported skill, performance features, task features. For each of these categories, there was an independent and a hierarchical version, thus producing twelve models to compare. Fitting and model comparison were performed separately for "self" and "other".

We compared models using the WAIC score (Watanabe, 2010), an approximation for the out-of-samples predictive log density, $-\mathbb{E}_X \left[\log \mathbb{E}_\theta p(X|\theta)\right]$, the outer expected value being computed with respect to the real underlying distribution of the data, and the inner expected value being computed with respect to the posterior distribution over parameters, $p(\theta|X_1, X_2, ...X_N)$.

Assuming a model with parameters θ and data points $\{X_1, X_2, ...X_N\}$, independent conditioned on parameters, WAIC is defined as

$$\begin{aligned} WAIC = &- \frac{1}{N}\Sigma_{i=1}^{N} \log(\mathbb{E}_\theta p(X_i|\theta)) \\ &+ \frac{1}{N}\Sigma_{i=1}^{N} \left[\mathrm{Var}_\theta(\log p(X_i|\theta))\right], \end{aligned} \tag{4.2}$$

where the expectations are taken with respect to the posterior distribution over the parameters. Given a set of samples from this posterior distribution, $\left\{\theta^1, \theta^2 ... \theta^S\right\}$, an estimator for WAIC can be computed using the sample averages of the quantities involved:

$$\begin{aligned} \widehat{WAIC} = &- \frac{1}{N}\Sigma_{i=1}^{N} \log(\frac{1}{S}\Sigma_{s=1}^{S} p(X_i|\theta^s)) \\ &+ \frac{1}{n}\Sigma_{i=1}^{N} \left[\mathrm{Var}_{s=1}^{S}(\log p(X_i|\theta^s))\right], \end{aligned} \tag{4.3}$$

For comparison with other commonly used model scores, Gelman et al (Gelman et al., 2014) recommend using a rescaled version of WAIC, denoted in the following by WAIC_G, which corresponds to not dividing by the number of data points and multiplying by 2. This is the version we used.

We obtained samples from the posterior distributions using the pystan interface (`https://pystan.readthedocs.io/en/latest/`) to the STAN probabilistic programming language (`https://mc-stan.org/`). For all the models we compared, we fitted a hierarchical and a non hierarchical implementation. In the non hierarchical implementation each subject was fitted independently of the others, while in the hierarchical implementation the model was augmented with a set of "population parameters" determining a Gaussian population distribution, from which we assumed individual subject parameters were drawn. This population distribution therefore acts as a prior for the individual subject parameters and its influence is reflected, alongside that of the data, in the posterior distribution over individual subjects parameters (Gelman et al., 2013).

Computing predictive likelihood, particularly in hierarchical models, im-

plies deciding on the relevant level of prediction and therefore of data grouping (Gelman et al., 2014). In both the hierarchical and the non-hierarchical setups, we were interested in the quality of prediction of new data from a subject given their individual parameters, and X_i stands for data from subject i.

All model dependent-analyses were performed separately for the "self" and "other" conditions.

4.3.2 Model comparison results

For both conditions, the model preferred by the WAIC scores was the full model, in its hierarchical version (see figure 4.22). This model was also preferred at the individual subjects level for most of the subjects: 23 out of 31 for the "self" condition and 16 out of 31 for the "other" condition. Of particular significance to us is that this indicates that previous skill responses do contribute to explaining attributions, over and above task features and performance features.

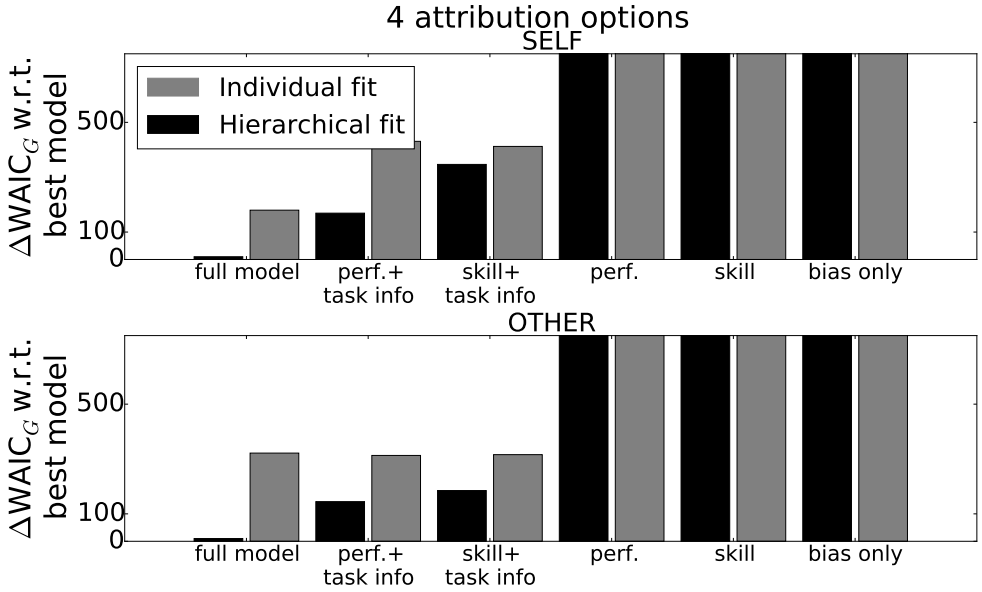


Figure 4.22: Model comparison, raw 4-options attributions. Bar heights show $10 + \Delta WAIC_G$ with respect to the best model. We shifted $\Delta WAIC_G$ by 10 for plotting purposes, to have a visible bar for the best model.

Figure 4.23 shows the match between real data and data obtained by

picking the most likely response label from the best model's prediction, for each subject. Analysis of the misclassifications revealed no interesting pattern: the model has a tendency to overpredict internal attributions and attributions to maze, which is not suprinsing, given the non-uniform distributions of subjects' responses for the different options, together with the fact that the cost of an error is the same, irrespective of the identities of the attribution options counfounded. More responses for the less preferred options are required for more detailed analyses of the errors, which could provide insights for more complex models. Task adaptation leading to more balanced attribution responses distributions is one of the goals for future work.

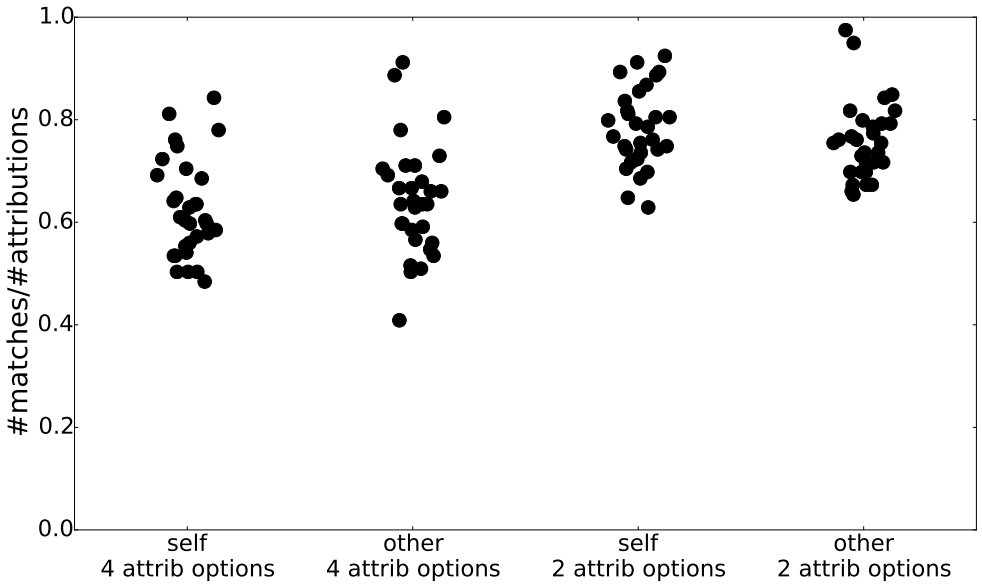


Figure 4.23: Proportion of matches between real responses and data generated greedily -by picking the most likely response- using the mean posterior parameters of the winning model. Each dot represents a subject, the x-axis coordinates have been jittered for plotting purposes. Left: accuracy on raw data. Right: accuracy computed for relabelling attributions as "internal" vs "external".

While the best model does a relatively good job of capturing the data, there is still room for improvement. The models we considered were all linear in the selected features, and adding interactions might improve them. Another avenue for more complex models is adding higher order effects of individual features.

One core goal of these analyses has been to establish whether previous skill responses contribute to attribution responses. The model comparison results showed that this is indeed the case. Further work, ideally on more - and more balanced - data, will allow for more in-depth investigation of subjects' attribution mechanism.

4.3.3 Model parameters

In this section we present analyses of the posterior parameters obtained from the best model (see figure 4.24 for the distributions of mean posterior parameters, across subjects [7]). The structure of the section mirrors that of the

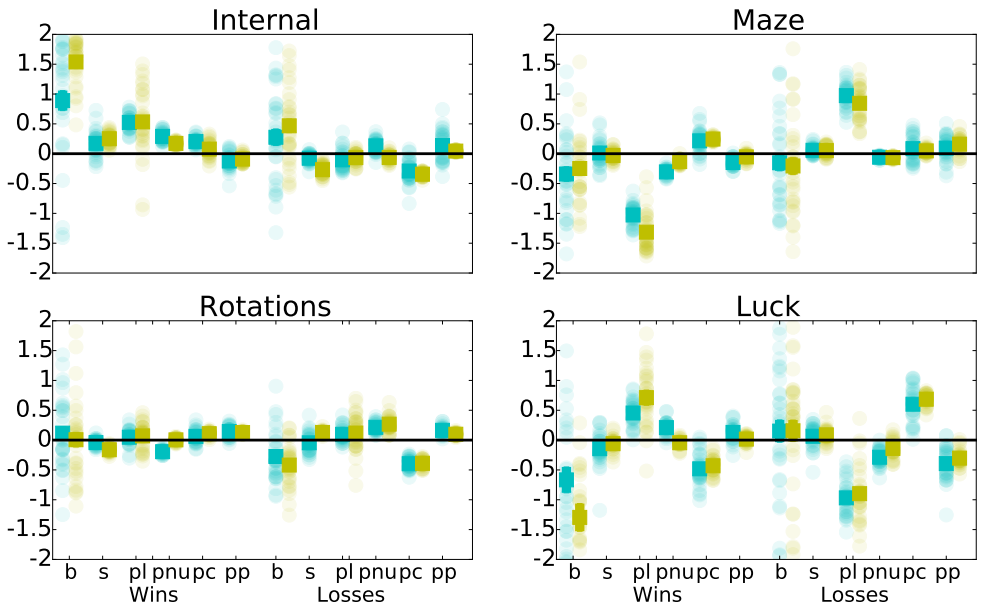


Figure 4.24: Distribution of mean posterior parameters, from best model for attribution responses. Parameters are grouped according to the attribution option. Only parameters for internal, maze and rotation options were independent parameters in the model, but we include the resulting parameters for attributions to luck for visualisation purposes. Faded dots represent individual subjects, thick lines show the mean $\pm$ s.e.m over subjects; "self"- blue, 'other'- yellow. In each axis, parameters for wins are grouped on the left and parameters for losses are grouped on the right. In all cases the order of the features is the same: bias, skill, path length, proportion of non up orientations, proportion of correct key presses, proportion of pauses post rotations. See appendix R for a detailed description of the model and the meaning of each parameter.

[7] Note that in all plots in this section, parameter names in figures have been reduced to the parameter name index, for convenience. Thus α_x, the weight of feature x, is labelled as x.

model-agnostic analyses: we present the effect of outcome, the effect of objective task and performance features and the effect of skill. As in previous analyses, parameters related to the contribution of previous skill responses to attribution, and comparisons between "self" and "other" are of particular interest, being directly related to our main questions. All reported significant effects survive the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) for correcting for multiple comparisons, unless otherwise stated.

A further matter of interest is whether the influence of skill on attributions is related to subject's responses to questionnaires. Analyses of the relationship between parameters and questionnaire measures are presented in the dedicated chapter (5).

4.3.3.1 Outcome

We tested for the presence of any self-enhancing or other-enhancing biases by comparing the model's estimates of the general preference that subjects show for making internal attributions after wins vs losses, in the absence of any additional information. This involves comparing the respective bias parameters in our model. However, because our model includes softmax transformations, the magnitude of the bias parameter for any given option does not directly and independently translate into the subject's preference for that option. It is the relationship between biases for the different options, entering into the softmax function, that determines preferences for the different options. Since there are separate softmax transformations for wins and losses, directly comparing the magnitude of parameters capturing the bias toward a given attribution option after wins vs losses is not meaningful. Instead, we applied the softmax transformations to the model reduced to bias parameters only, clamping all feature weights to 0, and compared the resulting probabilities. Table 4.1 shows the results of comparisons, done by repeated measures t-tests.

Across subjects, probabilities for making internal attributions for wins were significantly higher than their counterparts for losses for both "self" and "other". Consistent with our model-agnostic analyses results, we found that

	statistic	p-value	effect size
Self	2.72 (4.61)	0.01 ($7 * 10^{-5}$)	0.78 (1.2)
Other	7.5	$2 * 10^{-8}$	1.83

Table 4.1: Effect of outcome on bias toward internal attribution. Repeated measures t-tests results: probability of making internal attributions for wins vs for losses. Effect size computed as Hedge's corrected Cohen d. Bracketed values for 'self' were computed excluding the two subjects who provided no internal attributions for wins in the 'self' condition. We did not exclude these subjects from computations for 'other', as their responses in this condition were not severely different from those of other subjects.

the effect of outcome was stronger for "other" than for "self". Excluding the two subjects who provided no internal attributions for wins in the "self" condition increased the outcome effect for "self", but it still remained lower than for "other" (see table 4.1).

4.3.3.2 Objective task and performance measures

In our model-agnostic analyses we found several patterns of influence of the objective task measures and objective performance measures on attributions (see section 4.2.3). These analyses were performed separately for each variable of interest, and responses were averaged over trials corresponding to discretised levels of the variable of interest. In contrast, posterior model parameters were obtained from trial-by-trial modelling which included all variables of interest at once. It is therefore important to determine whether the patterns we have observed in model-agnostic analyses are born out at the level of model parameters as well. As we are about to show, this is largely the case. Additionally, parameter analyses also reveal effects that the coarser model-agnostic analyses could not detect, as we detail below.

The following analyses were restricted to external attributions to "maze" and "rotations", because we did not model weights for attributions to "luck" as independent parameters[8].

[8]We note that we present plots and results for the raw values of the weight parameters in this section: that is, the contribution of each feature to the score for a given attribution, before passing through the softmax; we also estimated the effects of each feature in the final space of probabilities, post softmax (see chapter 5) and obtained generally consistent results.

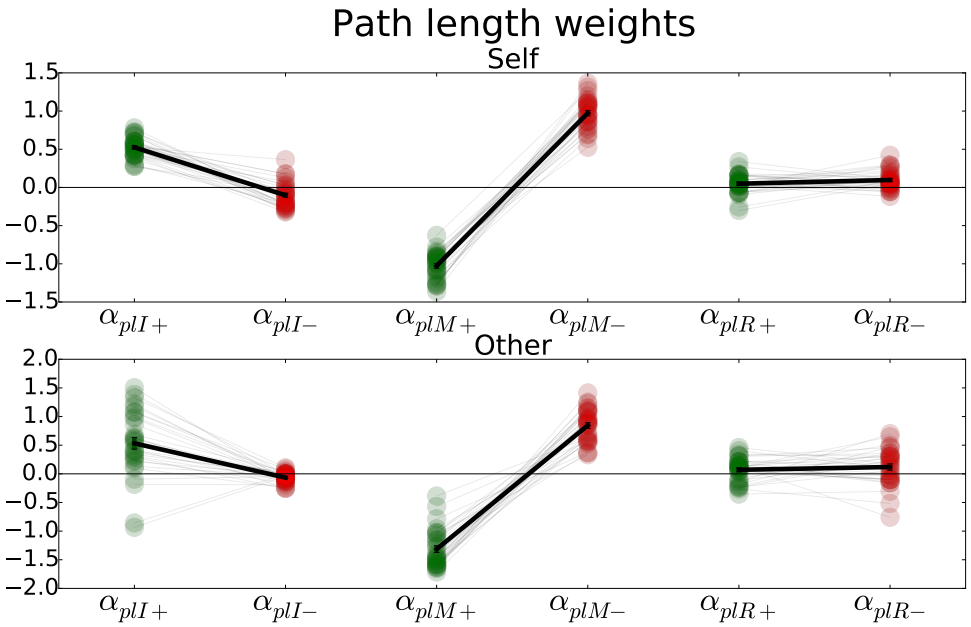


Figure 4.25: Mean posterior weights of path length for Internal, Maze and Rotation attribution options. Each dot represents a subject. Green and red represent wins and losses respectively. Heavy line represents the mean across subjects ± s.e.m. See appendix R for a detailed description of the model and the meaning of each parameter.

As far as path length is concerned, we found that for both "self" and "other" weights of this feature were positive for internal attributions after wins ("self": 1 sample $t(30) = 22.2, p < 0.01$, "other": 1 sample $t(30) = 5.12, p < 0.01$) and negative for internal attributions after losses ("self": 1 sample $t(30) = -3.45, p < 0.01$, "other": 1 sample $t(30) = -4.03, p < 0.01$), and the reverse for attributions to maze (wins: "self": 1 sample $t(30) = -33.24, p < 0.01$, "other": 1 sample $t(30) = -22.15, p < 0.01$, losses: "self": 1 sample $t(30) = 27.21, p < 0.01$, "other": 1 sample $t(30) = 16.78, p < 0.01$), see figure 4.25. Path length

Therefore increased path length increases the likelihood of making internal attributions for wins and decreases the likelihood of making internal attributions for losses, and produces the opposite effect on attributions to maze. This is to be expected as a reasonable way of using path length information and it is also consistent with what we found in our model-agnostic analyses.

Parameters associated with the proportion of non up orientations (see fig- Orientation

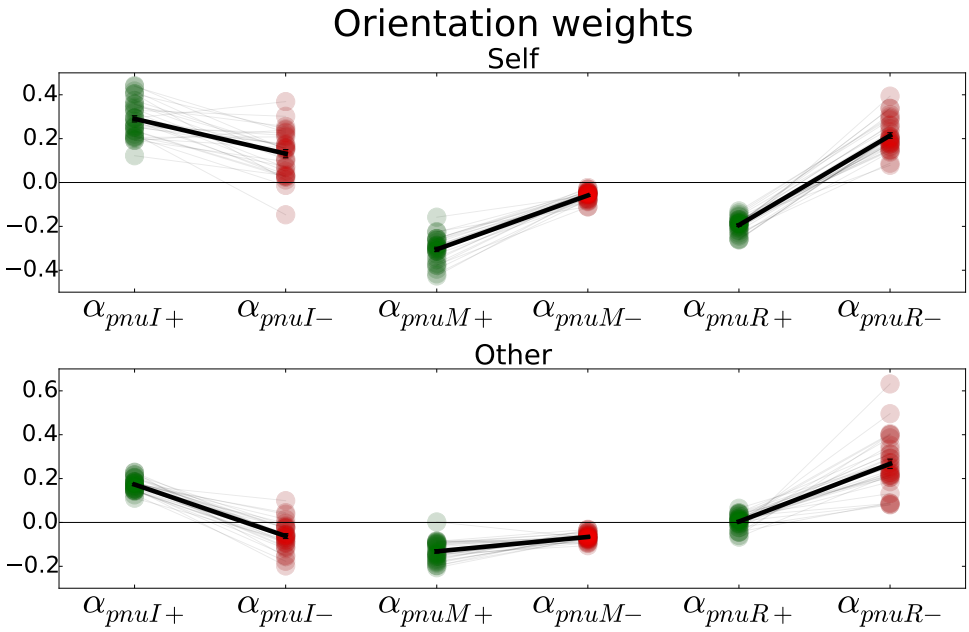


Figure 4.26: Mean posterior weights of the proportion of non up orientation for Internal, Maze and Rotation attribution options. Each dot represents a subject. Green and red represent wins and losses respectively. Think line represents the mean across subjects ± s.e.m. See appendix R for a detailed description of the model and the meaning of each parameter.

ure 4.26), which we included as a measure of the contribution of rotations to the difficulty of the trial, show both expected and unexpected characteristics, also recapitulating our model-agnostic observations (see section 4.2.3.2). We expected that increasing unusual orientations would increase the likelihood of internal attributions for wins and decrease the likelihood of internal attributions for losses, and that it would have the opposite effect on attributions to rotations. In terms of model parameters, this corresponds to positive values for α_{pnuI+} and α_{pnuR-} and negative values for α_{pnuI-} and α_{pnuR+}.

What we found only partially matched these expectations, but was consistent with our model-agnostic observations. Specifically, for the "self" condition our predictions were confirmed (win internal: 1 sample $t(30) = 19.71, p < 0.01$, win rotations: 1 sample $t(30) = -31.19, p < 0.01$, loss rotations: 1 sample $t(30) = 16.2, p < 0.01$), with the exception of the weights for internal attribution for losses. These were, contrary to our expectation but consistent with the ANOVA results in section 4.2.3.2, significantly larger

than 0 (1 sample $t(30) = 6.82, p < 0.01$), indicating that experiencing more unusual orientations increases the likelihood of internal attributions for losses, all else being equal.

For the "other" condition, our predictions were confirmed (win internal: 1 sample $t(30) = 33.07, p < 0.01$, loss internal: 1 sample $t(30) = -5.36, p < 0.01$, loss rotations: 1 sample $t(30) = 12.13, p < 0.01$), with the exception of the weights for attributions to rotations post wins, which were not significantly different from 0.

Parameters associated with performance (proportion of correct key presses) also confirmed the predictions we had concerning internal attributions (see figure 4.27), indicating that subjects monitored their and the "other"'s performance and used this information in their attributions. Thus, for both conditions, α_{pcI+} was significantly larger than 0 ("self": 1 sample $t(30) = 16.35, p < 0.01$, "other": 1 sample $t(30) = 3.39, p < 0.01$) and α_{pcI-} significantly lower than 0 ("self": 1 sample $t(30) = -8.2, p < 0.01$, "other": 1 sample $t(30) = -48.62, p < 0.01$), which means that increasing numbers of correct key presses are associated with increasing likelihood of internal attributions for wins and decreasing likelihood of internal attributions for losses. This is indeed consistent with our prediction that better performance would increase the likelihood of subjects taking credit for wins and avoiding blame for losses. Performance

Analyses of the weights for external attributions revealed surprising, as well as expected patterns. The likelihood of attributing losses to the maze does increase with better performance ("self": 1 sample $t(30) = 2.28, p = 0.03$, "other": 1 sample $t(30) = 5.41, p < 0.01$), however, surprisingly, so does the likelihood of attributing wins to maze: for both conditions, the parameters capturing the contribution of the proportion of correct key presses to attributions to maze after wins (α_{pcM+}) were significantly greater than 0 ("self": 1 sample $t(30) = 6.23, p < 0.01$, "other": 1 sample $t(30) = 41.96, p < 0.01$). indicating that subjects to some extent interpreted good performance as a re-

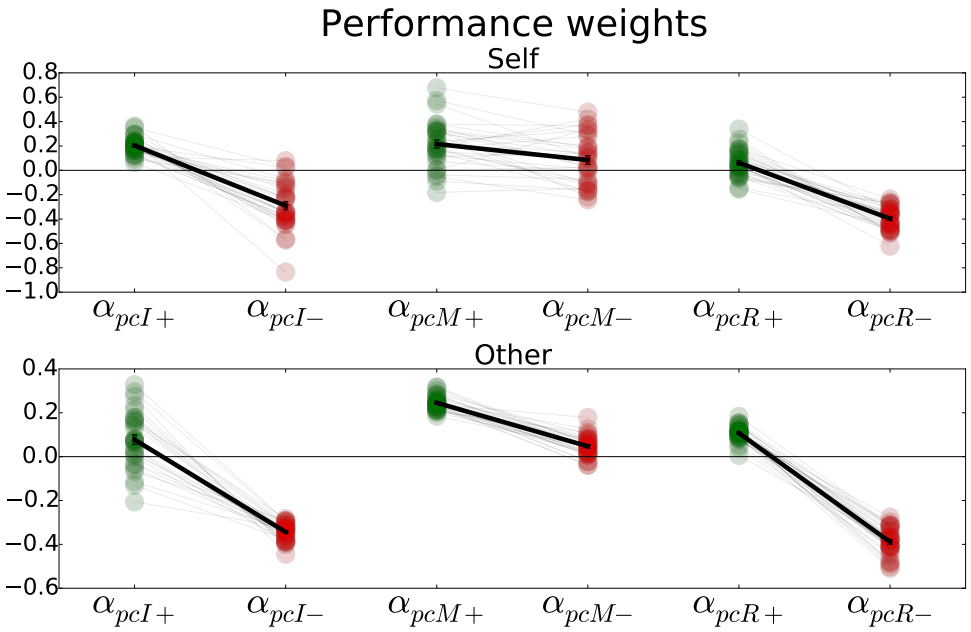


Figure 4.27: Mean posterior weights of the proportion correct key presses for Internal, Maze and Rotation attribution options. Each dot represents a subject. Green and red represent wins and losses respectively. Heavy line represents the mean across subjects ± s.e.m. See appendix R for a detailed description of the model and the meaning of each parameter.

flection of the maze being easy, rather than taking credit for it (or giving credit to the "other").

As far as attributions to rotations are concerned, for both "self" and "other" weights associated with attributions to rotations for wins are positive ("self": 1 sample $t(30) = 3.09, p < 0.01$, "other": 1 sample $t(30) = 17.17, p < 0.01$), while their counterparts for losses are negative ("self": 1 sample $t(30) = -24.02, p < 0.01$, "other": 1 sample $t(30) = -34.63, p < 0.01$). Thus, better performance is associated with an increase in the likelihood of attributing wins to rotations, and a decrease in the likelihood of blaming rotations for losses. Therefore, just as in the case of parameters for attributions to maze, subjects seem to credit the task for being easy when winning. Unlike the case of attributions to maze, however, better performance pushes blame for losses away from rotations. This might appear counterintuitive, but it is not unreasonable: since any contribution of rotations to losses happens through impaired performance, the better the performance, the less

rotations can be to blame for losses.

4.3.3.3 Previous skill report and comparison with performance

Parameters modelling the influence of previous skill reports on attributions were of particular interest to us, as they were directly related to our main question about the relationship between beliefs about self and causal attributions.

Model comparisons showed that previous skill reports contribute to explaining attributions responses over and above objective task and performance features, and further tests confirmed that with very few exceptions, weights of skill are significantly different from zero. Table 4.2 shows the detailed results of the one samples t-tests we performed.

		α_{sI+}	α_{sI-}	α_{sM+}	α_{sM-}	α_{sR+}	α_{sR-}
Self	statistic	4.64	-5.8	0.31	4.47	-4.07	-1.5
	p	< 0.01	< 0.01	0.76	< 0.01	< 0.01	0.14
Other	statistic	14.43	-16.31	-1.88	3.55	-20.28	23.75
	p	< 0.01	< 0.01	0.07	< 0.01	< 0.01	< 0.01

Table 4.2: One sample t-tests results: mean posterior parameters representing the weight of previous skill report vs 0.

We expected subjects to be more likely to assume responsibility for wins, and less likely to assume responsibility for losses, with increasing skill. This is indeed what we found(see figure 4.28): skill weights for internal attributions for wins were significantly larger than 0 and skill weights for internal attributions for losses were significantly lower than 0 for both conditions. This pattern is the same as the one we expected and observed for the effect of performance.

However the patterns for the parameters associated with external attributions are different. Weights of the proportion of correct key presses for attributions to maze showed that to some extent subjects explained their good performance in terms of the maze being easy, rather than taking credit for it (or giving credit to the "other"). This is not the case for skill, where α_{sM-} are significantly larger than 0, as expected, and α_{sM+} are not significantly dif-

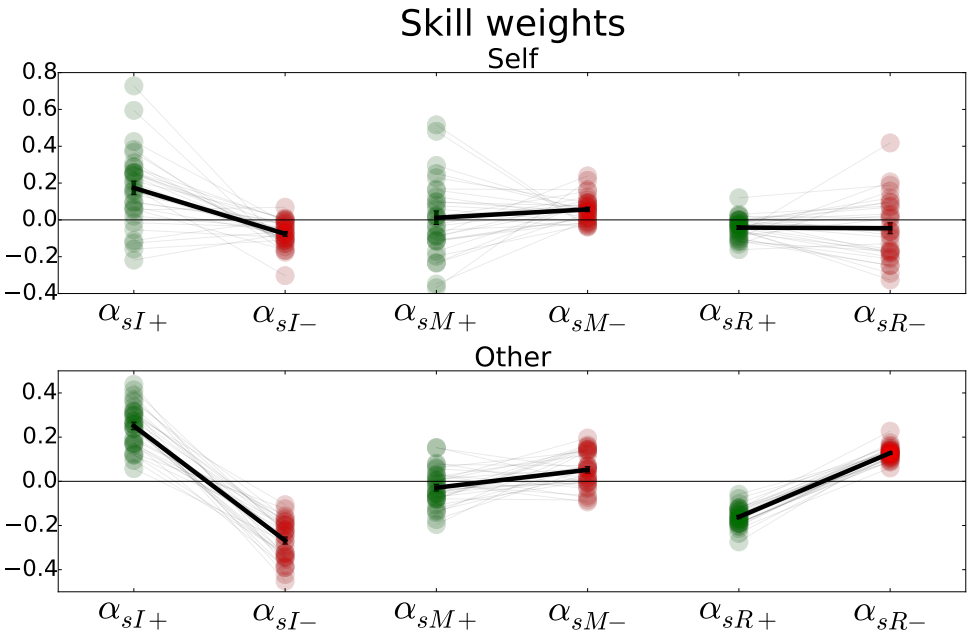


Figure 4.28: Mean posterior weights of the previous skill response for Internal, Maze and Rotation attribution options. Each dot represents a subject. Green and red represent wins and losses respectively. Heavy line represents the mean across subjects ± s.e.m. See appendix R for a detailed description of the model and the meaning of each parameter.

ferent from 0. This suggests that unlike momentary performance, skill is less likely to be explained away as an effect of maze simplicity.

The weights for attributions to rotations also reveal differences between the effect of performance and that of skill. The pattern of weights for skill is different from that of performance, and it is different in different ways for "self" and "other". Specifically, for "self", skill weights for attributions to rotations are negative for wins and negative, but not significantly so for losses, indicating that as skill increases, the likelihood of attributions to rotations decreases, irrespective of outcome. This is consistent with what skill means in the context of our task, where subjects have to learn to deal with rotations: as skill increases, rotations matter less.

For "other", the pattern is exactly the opposite of the one found for the parameters related to performance: higher skill is associated with decreased likelihood of crediting rotations for wins, and increased likelihood of blaming rotations for losses. Therefore for "other", but not for "self", the effect of

skill on attributions to rotations is similar to the effect on attributions to maze (see figure 4.28). This might be related to differences between watching and playing the task: it is likely that the difference between the perceived effect of rotations on performance when watching versus when actually playing is much larger than the difference between the perceived effect of maze structure in the two conditions.

These comparisons identified both expected and unexpected effects of skill on attributions. They also revealed differences between the effect of trial by trial performance (as measured by the proportion of correct key presses) on attributions and the effect of previous skill responses, differences that previous model agnostic analyses were unable to detect. These differences are consistent with a dissociation between a moment-by-moment measure of performance, which is vulnerable to explaining away in terms of external task features, and a measure of the less volatile underlying ability. This dissociation provides additional evidence for a specific effect of skill on attribution.

4.3.4 Model dependent analyses: summary

In this section we presented model-dependent analyses, investigating the effects of the factors of interest based on individual level, trial by trial prediction of subjects' attribution responses.

We compared classification models with a simple common structure, predicting attributions based on linear combinations of features. The preferred model, both across subjects and at the individual level, was the full model including all features - objective task and performance, as well as subjects' skill estimates. Thus subjects' skill estimates contributed to predicting their attributions over and above task and performance measures.

Analyses of posterior parameters from the best model generally confirmed observations from model-agnostic analyses, and in some cases they also provided additional insights. Note that while results of model-agnostic analyses involved separately testing for the effects of the factors of interest (with the exception of interactions with outcomes), the winning model ac-

counted for all factors simultaneously, albeit only linearly.

Consistent with model-agnostic analyses, we found that subjects displayed a moderate self-serving bias, along with a stronger other-serving bias.

The distributions of parameters related to objective task features revealed the expected effects for path length, and both expected and unexpected ones for orientations, consistent with observations from model-agnostic analyses.

For both "self" and "other", increasing path length was associated with increased likelihood of attributing wins internally and decreased likelihood of attributing losses internally, with the opposite patterns for attributions to maze (see figure 4.29).

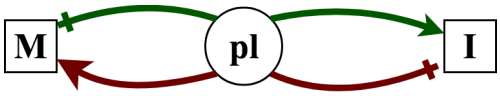


Figure 4.29: Weights of the path length (pl) feature on attributions to the maze (M) and internal attributions (I). Green arrows: weights for wins; red arrows: weights for losses. Pointed arrow heads: positive weights, blunt arrow heads: negative weights. Weight signs were the same for both conditions.

Parameters associated with orientations showed a similar pattern for the "other" condition, but not for the "self" condition (see figure 4.30).

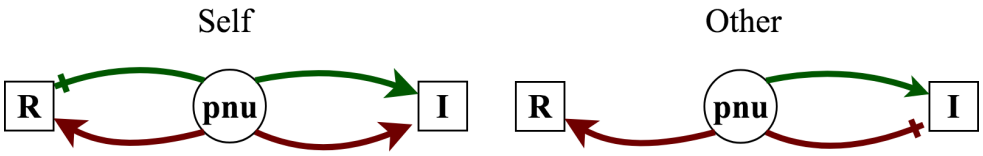


Figure 4.30: Weights of the proportion of non up orientations (pnu) feature on attributions to rotations (R) and internal attributions (I). Green arrows: weights for wins; red arrows: weights for losses. Pointed arrow heads: positive weights, blunt arrow heads: negative weights. Missing arrows indicate weights not significantly different from 0. Left: parameters for 'self'; right: parameters for 'other'.

Weights associated with internal attributions indicate that in the "other" condition subjects treated this rotation-related feature similar to the way the treated path length. However they processed it differently in the "self" condition, where an increase in the proportion of non up orientations was associated

with more internally assumed responsibility for both wins and losses, consistent with what we found in model-agnostic analyses. As we mentioned in our discussion of this observation in the model-agnostic section, this difference in the processing of the two features might reflect meaningful differences between their roles in the task. Subjects have direct control over the extent to which orientation changes impact performance, since correct adaptation of key presses can neutralise the effect of rotations. However they have less direct control over the impact of the maze structure on their performance. It is therefore not unreasonable that they would internalise responsibility for rotations more than they do for path length.

As far as performance is concerned (see figure 4.31), parameters for internal attributions showed the expected effects, increased performance being associated with increased likelihood of internal attributions for wins and decreased likelihood of internal attribution for losses in both conditions. However distributions of parameters for external attributions showed surprising patterns, suggesting that increased performance increased the likelihood of subjects crediting the task with being easy, while at the same time decreasing likelihood of blaming rotations for losses.

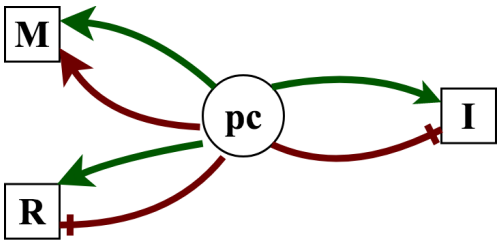


Figure 4.31: Weights of the proportion of correct key presses (pc) on attributions to the maze (M), rotations (R) and internal attributions (I). Green arrows: weights for wins; red arrows: weights for losses. Pointed arrow heads: positive weights, blunt arrow heads: negative weights. Weight signs were the same for both conditions.

This was not the case for parameters associated with skill estimates (see figure 4.32). In this case distributions of parameters for internal attributions confirmed our expectations and results of model-agnostic analyses. However distributions of parameters for external attributions showed a different pat-

tern from the one associated with performance parameters. Differences were consistent with a dissociation between a moment-by-moment measure of performance, vulnerable to explaining away in terms of external task features, and a less vulnerable measure of a more stable underlying ability. This dissociation provides additional evidence for a specific effect of skill on attribution and is consistent with results of model agnostic analyses indicating an effect of skill over and above performance.

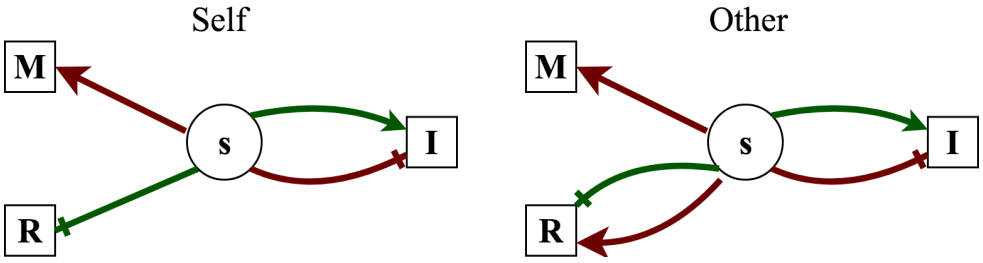


Figure 4.32: Weights of the skill (s) feature on attributions to the maze (M), rotations (R) and internal attributions (I). Green arrows: weights for wins; red arrows: weights for losses. Pointed arrow heads: positive weights, blunt arrow heads: negative weights. Missing arrows indicate weights not significantly different from 0. Left: parameters for 'self'; right: parameters for 'other'.

4.4 Reaction times

In this section we present analyses of subjects' reaction times when making attributions. We tested for any effect of condition ("self" vs "other"), and, separately for "self" and "other", for main effects of outcome and attribution ("internal" vs "external") and their interaction, as well as for secondary effects of skill.

Reaction times were significantly lower for "self" than for "other" (paired $t(15) = -4.06, p = 3 * 10^{-4}$), which is consistent with what we observed for skill responses reaction times, and with differences between responding for self vs other which have been documented in the past (Jackson et al., 2006; Kuiper and Rogers, 1979; Nowicka et al., 2018).

Following the same methodology that we used in our model-agnostic analyses of attribution responses (see section 4.2.3.2 and appendix O), we

performed permutation tests for main effects of outcome and attribution, as well as their interaction, separately for "self" and "other". We used the F-statistics employed in repeated measures two-way ANOVA, with outcome and attribution as fixed factors and subjects as random factors.

For both conditions, we found no effect of outcome, but a significant effect of attribution ("self" $F_{(1,30)} = 6.5$, p-value from permutation test, corrected for multiple comparisons $p = 0.03$, "other" $F_{(1,30)} = 21.38, p = 0$) and significant interaction ("self" $F_{(1,30)} = 6.95, p = 0.03$, "other" $F_{(1,30)} = 13.21, p = 0$). In both conditions subjects were faster in making external attributions. When making internal attributions, they were faster for wins than for losses, with the opposite pattern for external attribution. See figure 4.33.

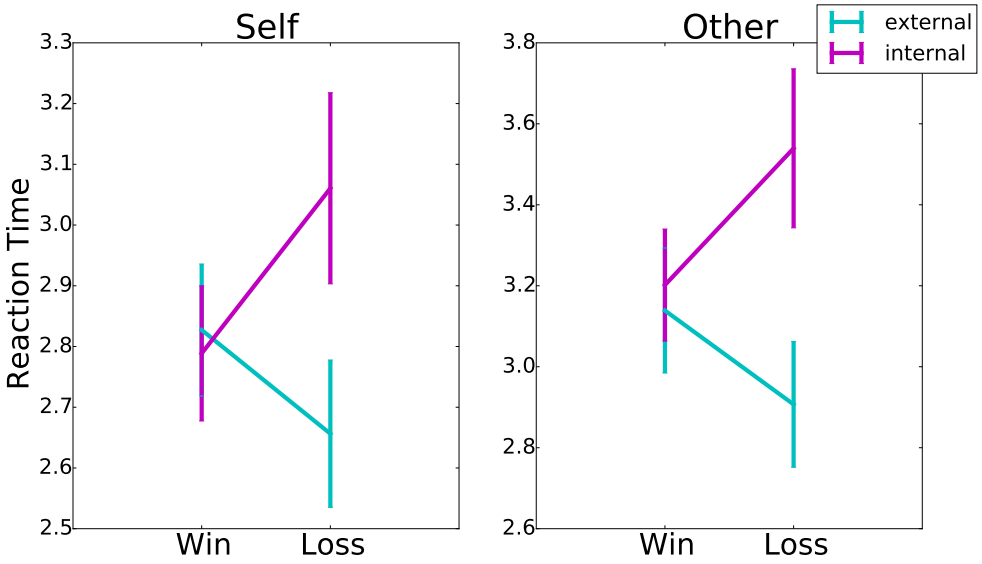


Figure 4.33: Attribution reaction times, effect of outcome, attribution and interaction. Mean and s.e.m across subjects. The value for each (outcome,attribution) pair for each subject was computed by averaging RTs from all relevant trials.

We expected that skill estimates would have an effect on RTs, conditioned on outcome and attribution, such that, for instance, higher skill would be associated with faster responses than slower skill, when making internal attributions for wins. In order to test for this, we regressed RTs on skill for each subject, separately for each condition and each (outcome, attribution) pair of possible values. We then tested whether the means of the resulting

distributions of skill weights across subjects were different from 0.

We found that skill had a significant effect on external attributions for wins in the "self" condition (1 sample $t(30) = -2.78$, p-value corrected for multiple comparisons $p = 0.04$) and a significant effect on internal attributions for wins in the "other" condition (1 sample $t(30) = -4.3, p = 0.001$).

The effect for "other" was as expected, namely skill was negatively correlated with RTs for internal attributions for wins, such that the better subjects reported the "other" to be, the faster they were in making internal attributions for wins. However in the "self" condition, the effect on RTs for external attribution was surprising, in that skill was also negatively correlated with RTs, indicating that the higher the skill, the faster subjects were in making external attributions for wins.

4.5 Summary

In this chapter we presented both model-agnostic and model dependent analyses of subjects' attribution responses.

Model agnostic analyses showed that subjects had a preference for attributing positive outcomes internally and negative ones externally, preference revealed in the proportions of respective attributions, as well as in reaction times, and that this bias was stronger when evaluating the "other" than it was when evaluating the self. Model-dependent analyses confirmed this observation. However, as far as proportions of internal vs external attributions are concerned, model agnostic analyses further showed that this bias was not present from the beginning of the task: in both conditions subjects started from making more internal attributions for negative outcomes than for positive ones, and then switched their preferences. As our task was framed as a learning task, we cannot determine to what extent this phenomenon was due to subjects' perception of improving at the task, rather than to their expectations about getting better.

We provided subjects with different options for task-related external at-

tributions - "maze complexity" and "rotations" - and measured the relevant task features - length of correct path through the maze, proportion of unusual maze orientation within a trial - in order to identify wether subjects were sensitive to variations along these dimensions. We found that subjects did display rational use of the available information about task features: both model agnostic and model-dependent analyses indicated that they made more internal attributions for wins associated with higher vs lower levels of these markers of task difficulty.

We also found evidence suggesting subjects processed these two aspects of the task differently; this might reflect either the fact that rotations and maze complexity are not orthogonal dimensions, as rotations are embedded within the maze on each trial, or the fact that learning in the task allows improvement in the management of rotations, while there is less room for improvement as far as dealing with a complex maze is concerned.

We found that subjects were also sensitive to performance, which influenced their attribution responses as expected: high vs low performance was associated with more internal responsibility for wins and less internal responsibility for losses. These observations were consistent between model-agnostic and model dependent analyses.

In addition, model agnostic analyses suggested the presence of an effect of skill over and above that of performance. Model dependent analyses also indicated differences between parameters associated with performance and those associated with skill estimates, consistent with different effects of momentary performance and skill. See section 4.6.3 below for a summary of results concerning the effect of skill.

4.6 Discussion

The impact of the various events one experiences on everything from one's mood to one's future behaviour depends on various aspects of the context and one's momentary state, but also, crucially, on the way the event is interpreted,

on what is believed to have caused it, and what it is believed to reveal about the world, others, and ourselves. In particular, the extent to which one is responsible for the outcome is an essential part of the event appraisal, and one that is involved in a diverse range of phenomena, from questions of morality to aspects of psychological well-being, to practical decision making.

The way humans make causal attributions has therefore been a topic of interest for psychology, as well as psychiatry, and aspects of it have been extensively studied(see literature review in chapter 1). Among them we mention, since they are relevant to our discussion later on, the "actor-observer" effect, according to which people tend to interpret the behaviours of others as more indicative of internal traits, while reporting external causes to be more important in explaining their own behaviour (Jones and Nisbett, 1987), the self-serving bias prompting people to assume disproportionately more responsibility for positive outcomes than they do for negative ones (Alloy and Abramson, 1979; Tillman and Carver, 1980; Martin et al., 1984; Vázquez, 1987; Bentall and Kaney, 2005) (see Campbell and Sedikides, 1999; Mezulis et al., 2004, for reviews), along with its counterpart, the "depressive realism" ((Alloy and Abramson, 1979; Martin et al., 1984; Vázquez, 1987), and finally the association between a "negative attributional style" and depression (Peterson and Seligman, 1984; Lyon et al., 1999), (see Sweeney et al., 1986; Robins, 1988; Robins and Hayes, 1995, for reviews). There are, however, many aspects of even these phenomena which remain controversial and or poorly understood (Miller and Ross, 1975; Heine and Hamamura, 2007; Robins et al., 1996; Malle, 2006) .

Our interest in this study was, in particular, the relationship between attributions and beliefs about oneself, which might be a key ingredient in a better understanding of both attribution and its effect on psychiatric disorders (Bentall et al., 2001; Bentall, 2003). The main question we set out to address was whether subjects' beliefs about their own skill (as the relevant instantiation, in our task, of the more general "belief about the self") has an effect on

the way they make attributions for the outcomes they experience in the task. Additionally, we aimed to determine whether there are differences between any such effect in the case of making attributions when oneself is involved, vs making attributions when watching others (in our case the subjects watched their own replayed trials, under the pretence of watching some other subject's previously recorded performance; see chapter 2).

We begin by discussing the presence or absence in our data of the often observed effects mentioned above, then present a summary of our results concerning the relationships between skill and attributions, and our conclusions about the main questions of the study. We end this discussion by presenting some task improvements and directions and questions for future work, directly related to the attribution data. Given the novel, exploratory nature of the task, there are numerous design changes and questions that these data suggest, and we come back to them in the final section of this thesis (see 6.3).

4.6.1 Actor-observer effect

The actor-observer effect (Jones and Nisbett, 1987) (AOE) has been observed in experiments in which subjects were asked to perform a task and explain their behaviour (actor condition), as well as to watch another person perform the task and explain this other person's behaviour(observer condition). The effect refers to subjects' tendency to rely more on external, situational factors when explaining their own behaviour, and to rely more on factors internal to the actors, rather than situational constraints, when explaining the behaviour of the other person.

AOE has been investigated with a variety of tasks, such as ones in which subjects had to associate state or trait-like words with themselves and the actors (Nisbett et al., 1973), or tasks in which they were asked to evaluate progress in the learning of a novel task (Miller, 1975), or predict future performance (see Kelley and Michela, 1980; Malle, 2006, for reviews). Evidence has, however, not all been consistent, and particular aspects of task framing, design and implementation (such as focusing on trait inference vs explain-

ing behaviour, the coding of subjects' responses, the valence of the outcomes used, between vs within subject design, using hypothetical vs real outcomes) might have played a significant role in some of the studies establishing the AOE (Robins et al., 1996) (see Malle, 2006, for a review).

The relationship between the actor-observer effect and self-serving bias is also not straightforward. Malle's review (Malle, 2006) indeed concludes that, despite it being for a long while accepted as an established fact, there is insufficient evidence for a general AOE, and that while negative outcomes might be associated with the AOE, the reverse might be true for positive ones.

In our task, we found that subjects made significantly more internal attributions for "other" than they did for themselves, thus providing evidence in favour of the AOE. This is particularly interesting given that in this case subjects were not watching a real other person, but only a behaviour, and therefore the effect cannot have been produced by confounds such as the saliency of the actor in the observer condition. Other factors that might have contributed to the AOE effect in previous studies were also absent in our task: we used a within-subject design, attributions referred to real, not hypothetical behaviour, and subjects were presented with the same visual input in the two conditions.

Of note, the effect was significant for wins, but not for losses (internal attributions "self" vs "other" wins: paired $t(15) = 3.62, p = 0.001$, losses: paired $t(15) = 0.95, p = 0.35$). This model-agnostic observation is consistent with the result of model-dependent analyses: we compared preferences for internal attributions, computed from best model parameters, and found higher preferences for internal attributions in the "other" than in the "self" condition for both wins and losses, with the difference being significant for wins, but not for losses.

These results are also consistent with a general positivity bias (Tillman and Carver, 1980), and they indicate the presence of stronger other-serving biases than self-serving ones, which we discuss further in the next section.

4.6.2 Self-enhancement

People's tendency to adopt and attempt to maintain a positive view of themselves - by selectively attending to positive vs negative information or by overweighting positive information - has long been documented in a variety of tasks and contexts (Lyons et al., 2020) (see Blaine and Crocker, 1993; Campbell and Sedikides, 1999; Mezulis et al., 2004, for reviews).

Such self-serving biases can be argued to be beneficial, since they contribute to the maintenance of well being, as well as promoting persistence and exploration in the face of negative feedback or failure. Indeed associations between higher levels of such biases and external measures of success in very competitive environments have been documented (Lyons et al., 2020), as has their absence from depressed patients, an effect dubbed "depressive realism" (Alloy and Abramson, 1979; Martin et al., 1984; Vázquez, 1987).

On the other hand, it could equally well be argued that heightened attention to negative feedback, particularly in harsh environments, can be essential to survival, enabling animals to quickly learn from bad outcomes and avoid them in the future. There is indeed ample evidence of the privileged status of processing negative feedback, in animals as well as humans (Maier and Seligman, 2016; Müller-Pinzler et al., 2019). Negative, more than positive feedback, produces rapid and strong bodily responses, mobilising the organism for reaction; negative emotions produce more arousal than positive ones; negative events and information focus attention (see Taylor, 1991, for a review). In humans, concepts for negative actions and consequences form earlier than their positive counterparts (Fincham, 1985), negative events are surveyed more for potential causal information (Wong and Weiner, 1981; Bohner et al., 1988), and they elicit more spontaneous causal attributional activity than positive ones (Peeters and Czapinski, 1990).

Negativity bias - the tendency to overweight negative information - has been documented in social judgements (Müller-Pinzler et al., 2019), as has self-effacing (Akimoto and Sanbonmatsu, 1999; Deaux and Farris, 1977;

Heine and Hamamura, 2007), particularly in non-western cultures. Indeed the extent to which self-enhancing biases are a universal constant rather than a culturally specific phenomenon has been much debated (Mezulis et al., 2004; Heine and Hamamura, 2007).

"Other-serving" biases have also been documented, and people have been observed to be "nicer" to other than to themselves in past studies, not related to causal attributions (Crockett et al., 2014; Rand et al., 2014; Rand and Nowak, 2013).

We found evidence for both self-enhancement and "other"-enhancement in our data. The latter appeared stronger than the former, with subjects making more internal attributions for wins in the "other" condition than in the "self" condition, and showing a larger difference in the preference for internal attributions after wins vs after losses for "other" than for "self". Results of comparisons between parameters obtained through model-fitting were consistent with these model-agnostic observations. Probabilities computed based on bias parameters only revealed that the preference for making internal attributions for wins was significantly higher than its counterpart for losses in both conditions, and that this effect of outcome was stronger for "other" than for "self". We note that removing the two subjects who provided no internal attributions for wins in the "self" condition lead to an increase of the effect of outcome for "self", but the effect remained stronger for "other".

Analyses of reaction times also revealed biases for positive information: subjects were faster in making internal attributions for wins than they were in making internal attributions for losses, with the opposite pattern for external attributions. This effect was present in both conditions, but also stronger for "other" than for "self".

Our data therefore provides evidence for "self-enhancing" biases, and even stronger evidence for "other-enhancing" biases. This pattern was also present in analyses of the skill responses, where effects were generally larger for "other" than for "self", so it might be due to a generally higher level of

noise in the "self" condition.

The fact that there was no real other, and instead subjects watched their own previous performance complicates matters. Debriefing questionnaires responses showed that most subjects had some suspicions about the deceit, but we do not have a clear picture of the extent to which subjects recognised their own previous performance, and if they did, at what time in the task this happened. Due to the body of literature providing evidence for diminished or absent self-enhancement, as well as the presence of its opposite - self-effacement - in non-western cultures and particularly in East-Asian ones (Akimoto and Sanbonmatsu, 1999; Heine and Hamamura, 2007), and given that our subject population was drawn mostly from students at UCL, which has a large East-Asian student population (`https://www.ucl.ac.uk/srs/student-statistics`), it is unfortunate that we did not collect the relevant information to establish whether such cultural patterns might help explain the "other-enhancing" effect present in our data.

4.6.3 Effect of skill reports on attributions

Our main purpose was to investigate the following three questions: whether subjects' belief about their own skill at the task contribute to the way they make causal attributions for their outcomes; how this contribution compares with those of other relevant factors, such as outcome, objective measures of task difficulty and objective performance; and whether there are differences in the mechanisms through which belief about skill contributes to causal attributions when the self is involved, vs when attributions are provided for events involving another.

We found that skill does have an effect on attributions, as evidenced by the results of model-agnostic tests, as well as model comparison, and analysis of model parameters recovered from the best model. Model comparison preferred the full model, which in addition to objective task and performance measures included previous reported skill. Thus skill contributes to explaining attribution responses, over and above outcome and measurable task and

performance features. Both model-agnostic analyses and analyses of model-parameters showed that increasing skill is associated with increasing likelihood of internal attributions for wins and decreasing likelihood of internal attributions for losses.

Model-agnostic tests showed that the effect of skill on internal attributions persists when controlling for performance. Model-dependent analyses, which, being based on trial by trial modelling, can provide a finer-grained picture of the data, also identified differences between the effect of skill and that of moment-by-moment performance, specifically ones consistent with reported skill being a more stable measure of ability. Thus, while model-parameter analyses indicated that subjects to some extent explained away high moment-by-moment performance in terms of the task being easy, no such pattern was present for high skill.

These observations were valid for both conditions, but skill had different effects in the two conditions as far as attributions to "rotations" were concerned, pointing to differences between the way subjects perceive the effect of rotations and/or the way they conceptualise skill when they are the agents vs when they are merely watching.

4.6.4 Dynamic interactions between skill and attributions

The motivation for this work comes from Bentall's theory of interaction between attributions and beliefs about the self (Bentall et al., 2001; Bentall, 2003) . This theory posits reciprocal influences exist between these two variables, which, under normal circumstances, contribute to maintaining psychological well-being; however significant negative events can push the system out of this dynamically maintained balance and into catastrophic vicious circles, leading to psychiatric disorders.

We aimed for our task to provide time courses of attributions and belief about skill, in order to be able to study the loop connecting the two variables. In chapter 3 we have presented analyses aimed at identifying one of the arrows in the system - the influence that attributions exert on beliefs about the skill.

In the present chapter we have presented the opposing arrow - the effect of beliefs about skill on attributions. The results we presented in this chapter and in its counterpart dedicated to skill responses establish that effects in both directions can be detected within our task.

However this separation of the two mechanisms constitutes only the first steps towards the study of their potentially complex time-varying interactions. A major limitation of our analyses has been the assumption that relationships between these variables are stable in time and across levels of the two variables: we have not allowed for time varying effects, nor for potentially different regimes in different ranges of skill or attribution propensities. These limitations are due to the novelty of the task and the nature of this experiment, aimed chiefly at establishing a proof of concept and constitute a pilot for future research. Additional data, including data from patients, as well as task adaptation and more sophisticated analyses are needed to tackle such complexities. We provide a more detailed discussion of the directions for future work in the final chapter of this thesis (see 6.3).

4.6.5 Conclusions and future work

We conclude that subjects displayed reasonable integration of task and performance measures in their causal attributions, and that their own previous reports of skill contributed to explaining their responses, over and above other such features. This work therefore provides evidence that in the context of repeatedly experienced real outcomes, when subjects provide both causal attributions and skill evaluations, reported skill (a proxy for subjects' real underlying belief about their ability) has an effect on causal attributions.

Differences between "self" and "other" are also present. These generally involve stronger effects for "other" than for "self", and in particular stronger other-serving biases - biases towards positive information - however we cannot with the present data determine whether this is not merely a consequence of higher noise in the more emotionally salient "self" condition. Such self-other differences could also reflect cultural components, however our present

data is not suitable for investigating this hypothesis.

Due to the exploratory nature of the experiment, there are a number of features that were not optimised, and that future work would need to improve upon. We will come back to these in the more extensive discussion at the end of this thesis (see 6), but we mention here some of the aspects particularly relevant to attribution data.

The labelling of attribution response options was based on the need to provide both internal and external options, and as far as the external options were concerned, on the goal of providing options related to the different objectively measurable task features. However this introduced an unfortunate availability bias, which future work should remove. We also note that providing subjects with discrete options might be less informative than asking them to rank the potential explanations for the outcomes, or provide estimates of the extent to which they would assign responsibility to each.

Our choice of the task aspects to mention in the external attribution options was based on our hypotheses about which features of the task would be most relevant to subjects. However they included asymmetries in the natures of the external options provided, as the maze and rotations contributed differently to the task structure, with maze providing the context, and rotations being the aspect most directly controllable by subjects.

Finally, we chose to only ask for attribution responses and skill responses once every two trials, due to time constraints. However this had the undesirable effect of introducing gaps in our time-series of belief and attribution measures, without allowing for enough trials in-between measurements to allow for an averaging effect.

These aspects introduced undesirable complexities in both performing and interpreting analyses. However, despite these difficulties, we found convincing evidence of skill effects of attributions, as well as interesting differences between the processing of different task aspects and between self and other, which are worth investigating further, and provide hypotheses for more

targeted future research. See the final section of this thesis for further discussion.

Chapter 5

Questionnaire measures

In order to investigate relationships between behaviour in our task and well-established questionnaire-based measures of related psychological dimensions, we administered three questionnaires: the Rosenberg Self-Esteem Scale(SE) (Rosenberg, 1965), the Levenson Locus of Control Scale (LC) (Levenson, 1974) and the Attributional Style Questionnaire (ASQ) (Peterson et al., 1982)(see appendix E for the questionnaires as administered to our subjects). In this section we present analyses of the questionnaire responses, and of their relationships with behavioural data from our task and model parameters obtained in previously described model-dependent analyses (see 4.3). Given the small number of subjects in our dataset, we conceive of these analyses primarily as an explorative tool to generate hypotheses and define questions for future work.

The chapter is structured as follows: we begin with an overview of responses on each questionnaire, and present the dimensionality reduction approach we used in subsequent analyses; we then present analyses of relationships between questionnaire responses and descriptive statistics of behavioural data; we then present analyses of their relationships with model parameters obtained from the best model of attribution responses (see 4.3). The chapter ends with a discussion.

5.1 Questionnaire scores - overview

5.1.1 Self esteem

The Rosenberg Self Esteem Questionnaire (Rosenberg, 1965) is a ten-item scale, each item being a positive or negative statement about oneself. Subjects have to indicate the extent to which they agree with each of the statements, choosing between 4 available options (Strongly Agree, Agree, Disagree, Strongly Disagree, worth 4 to 1 points respectively; items containing negative statements are reverse scored). The score is obtained by summing the scores for all ten items. Higher scores indicate higher self-esteem.

We chose this scale because it has been extensively used and its consistency and external validation have been established (Schmitt and Allik, 2005). Scores obtained on this scale have been repeatedly and reliably found to be negatively correlated with neuroticism, positively correlated with extraversion, and positively correlated with positive self-models in the context of romantic attachment (Schmitt and Allik, 2005; Martín-Albo et al., 2007).

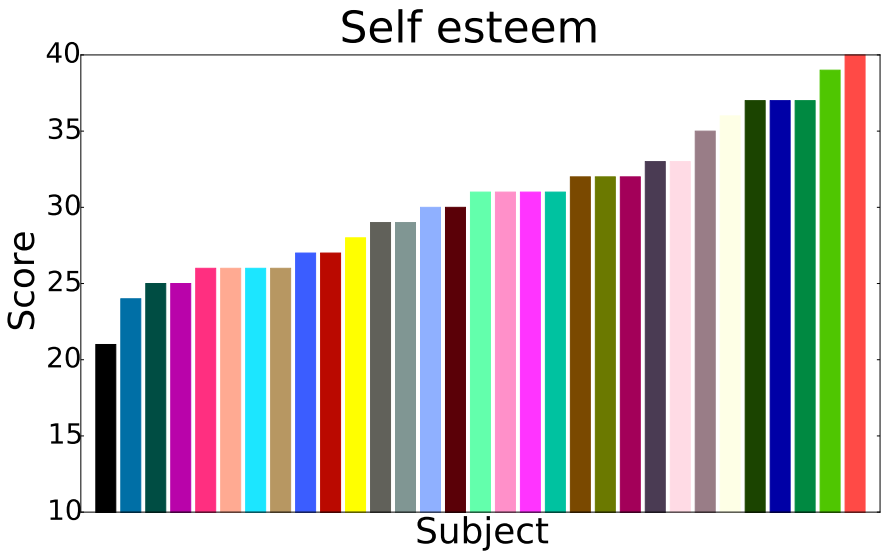


Figure 5.1: Scores on the Rosenberg self-esteem scale. Possible scores range from 10 to 40. Higher scores indicate higher self-esteem. Mean 30.52, s.d. 4.65. Colours indicate individual subjects, and correspond to colors used in figures 5.2 and 5.3.

Figure 5.1 shows the scores of our subjects. The mean and standard

deviation of our sample (mean = 30.52, s.d. = 4.65) are similar with statistics obtained for UK subjects in (Tafarodi and Walters, 1999) and Schmitt et al (Schmitt and Allik, 2005), an extensive study across cultures, which had a much larger population of respondents from the UK[1].

Our subjects' responses display a generally observed tendency toward positive self esteem, in that the mean value is significantly higher than the midpoint of the scale, 25 (1 sample $t(30) = 6.5, p = 3 * 10^{-7}$). This phenomenon has been repeatedly observed in samples across cultures (Schmitt and Allik, 2005).

5.1.2 Locus of Control

Levenson's locus of control questionnaire (Levenson, 1974) contains 24 items - statements about the control that oneself, others or chance have over events. Subjects have to indicate the extent to which they agree with each of the statements choosing one of 6 options (from "Strongly disagree", coded -3, to "Strongly agree", coded 3). Scores for each subscale are obtained by adding the responses for the eight relevant items and shifting the result by 24, to obtain positive values. The possible range is therefore 0-48 for each subscale. For each subscale, higher values indicate perception of higher levels of control exercised by the respective agent.

The three factor structure of this questionnaire has been confirmed repeatedly (Levenson, 1973, 1974; Walkey, 1979; Brosschot et al., 1994; Presson et al., 1997), and correlations with various psychological characteristics (Brosschot et al., 1994),subjective stress, neuroticisn (Morelli et al., 1979), depressive symptomatology (Moreira et al., 2020; Presson and Benassi, 1996; Presson et al., 1997), and coping mechanisms (Vickers Jr et al., 1983; Butler and Burr, 1980; Brosschot et al., 1994) have been found.

Figure 5.2 shows the scores of our subjects. Statistics of scores in our sample (Internal: mean = 35.65, s.d. = 4.64, Others: mean = 18.87, s.d. =

[1]Note that this is the case despite the fact that our subjects were recruited through mailing lists containing large numbers of students, many of them foreign, so our sample might not be representative for the UK to the extent that the previous study's sample was.

7.81, Chance: mean = 20.16, s.d. = 7.85) agree with previously published statistics for a variety of population samples (Hyman et al., 1991; Walkey, 1979; Levenson, 1974) and display the common pattern of significantly higher score on the internal subscale than on the two external subscales (Internal vs. Others[2] paired $t(15) = 9.24, p = 8 * 10^{-10}$, Hedge's corrected Cohen d = 2.54; Internal vs. Chance paired $t(15) = 7.71, p = 4 * 10^{-8}$, d= 2.33; Others vs. Chance paired $t(15) = -0.85, p = 0.4$, d = -0.16).

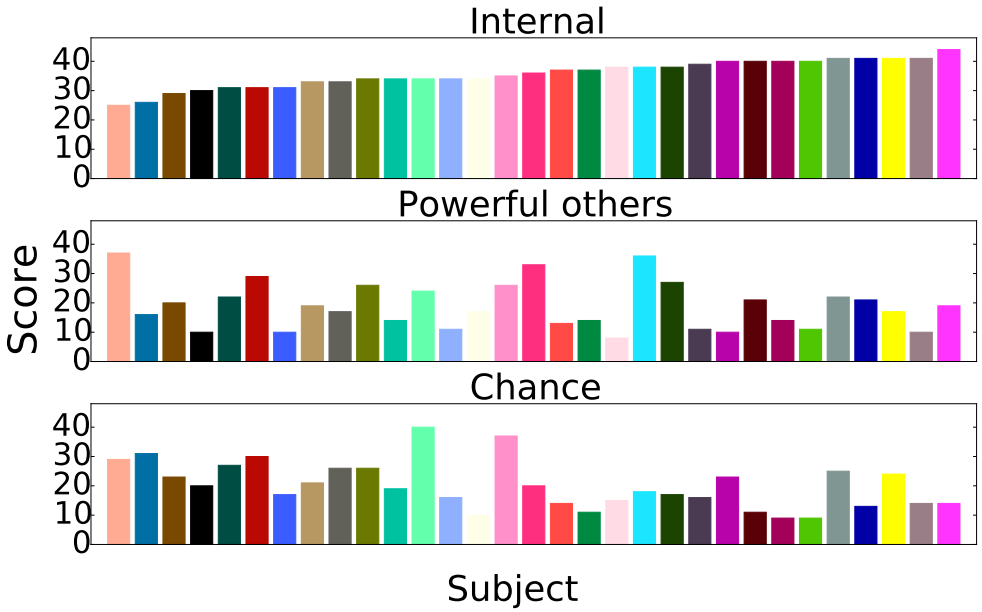


Figure 5.2: Scores on the Levenson locus of control scales. Possible scores range from 0 to 48. Higher scores indicate higher perceived control of the respective agent. Subjects ordered according to internal lc scores. Internal: mean = 35.65, s.d. = 4.64, Others: mean = 18.87, s.d. = 7.81, Chance: mean = 20.16, s.d. = 7.85. Colours indicate individual subjects, and correspond to colors used in figures 5.1 and 5.3.

5.1.3 Attributional Style

The Attributional Style Questionnaire (Peterson et al., 1982) includes 12 items. Each presents a situation and asks subjects to imagine themselves experiencing it and to report what they feel would be the major cause if that event happened to them. Subjects are then asked to rate, on a scale from 1

[2]P-value for results reported as significant have been Bonferroni-corrected (Bonferroni, 1936) for multiple comparisons.

to 7, the following four aspects: how important that event would be for them (importance), to what extent the cause was due to themselves or to "other people or circumstances" (internality), to what extent the cause would again be present in similar situations (stability) and to what extent the cause would affect other areas of their lives, beyond the described situation (globality). Items can be classified according to a number of subscales, but the ones relevant to us in this work are the six subscales corresponding to positive and negative internality, stability and globality. For each of these, the score is obtained by averaging the corresponding responses, thus obtaining scores ranging between 1 and 7.

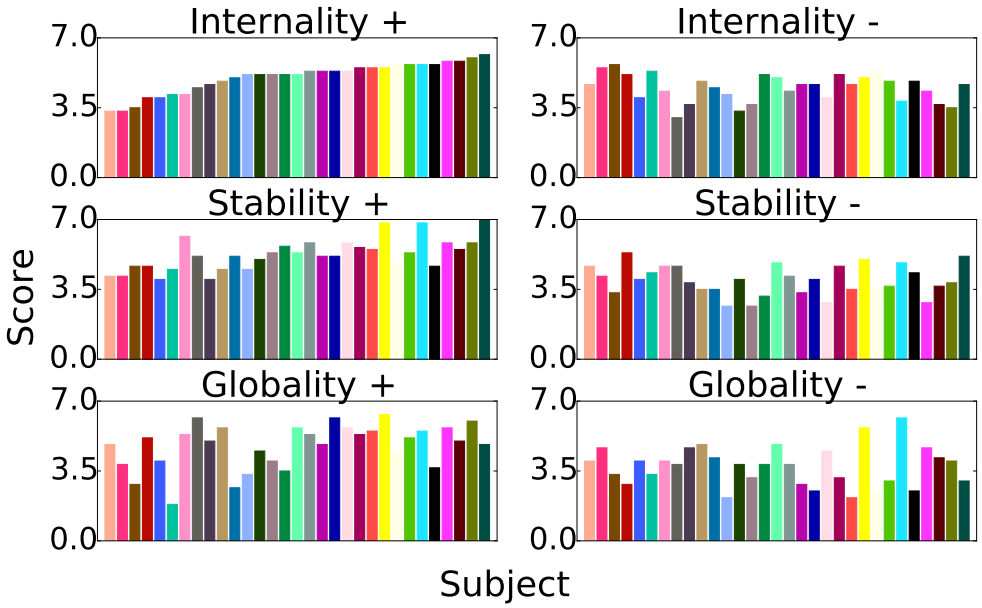


Figure 5.3: Scores on the ASQ scales. Possible scores range from 1 to 7. Higher scores indicate higher internality, stability and globality respectively. Subjects ordered according to internality for positive items. Internality+: mean = 5.02, s.d. = 0.77, Internality-: mean = 4.5, s.d. = 0.67, Stability+: mean = 5.25, s.d. = 0.79,Stability-: mean = 3.95, s.d. = 0.73, Globality+: mean = 4.77, s.d. = 1.1,Globality-: mean = 3.74, s.d. = 0.97. Colours indicate individual subjects, and correspond to colors used in figures 5.1 and 5.2.

The ASQ was developed as a tool to measure cognitive vulnerability, defined within the reformulated learned helplessness theory of depression (Abramson et al., 1978)(see also literature review in chapter 1) as the ten-

dency to attribute negative life events to internal, stable and global causes. The ASQ's psychometric characteristics have been widely studied (Peterson, 1991; Higgins et al., 1999; Hewitt et al., 2004) and both the ASQ and other questionnaires developed from it (Meins et al., 2012), such as the Cognitive Style Questionnaire (Haeffel et al., 2008), have been found to correlate with, or predict with various degrees of success aspects of depressive symptomatology (Alloy et al., 2000; Giuntoli et al., 2019) (see also literature review in chapter 1).

Figure 5.3 shows the scores for our subjects; summary statistics (Internality+: mean = 5.02, s.d. = 0.77; Internality-: mean = 4.5, s.d. = 0.67; Stability+: mean = 5.25, s.d. = 0.79; Stability-: mean = 3.95, s.d. = 0.73; Globality+: mean = 4.77, s.d. = 1.1; Globality-: mean = 3.74, s.d. = 0.97) are similar to the ones reported by Peterson et al (Peterson et al., 1982) in the original paper.

Consistent with the positive shift with respect to the middle point of the range that we saw in the self esteem scores, mean scores for the positive dimension were higher than the ones for the negative one for all subscales; the difference was significant for stability and globality subscales(Internality + vs. -: paired $t(15) = 2.5, p = 0.054$, d = 0.7; Stability + vs.-: paired $t(15) = 7.45, p = 8 * 10^{-8}$, d = 1.65; Globality + vs. -: paired $t(15) = 4.53, p = 2 * 10^{-4}$, d = 0.96). We do not know to what extent this is a general pattern in normal controls: while there has been interest in using the ASQ in relation with resilience (Needles and Abramson, 1990; Kleiman et al., 2013; Haeffel and Vargas, 2011; Johnson et al., 2017), the ASQ was designed and primarily used as an instrument to study aspects of depression and negative mood (see literature review in chapter 1) . Therefore negativity and scores on the negative subscales have generally been the focus of interest, rather than relationships between negative and positive subscales.

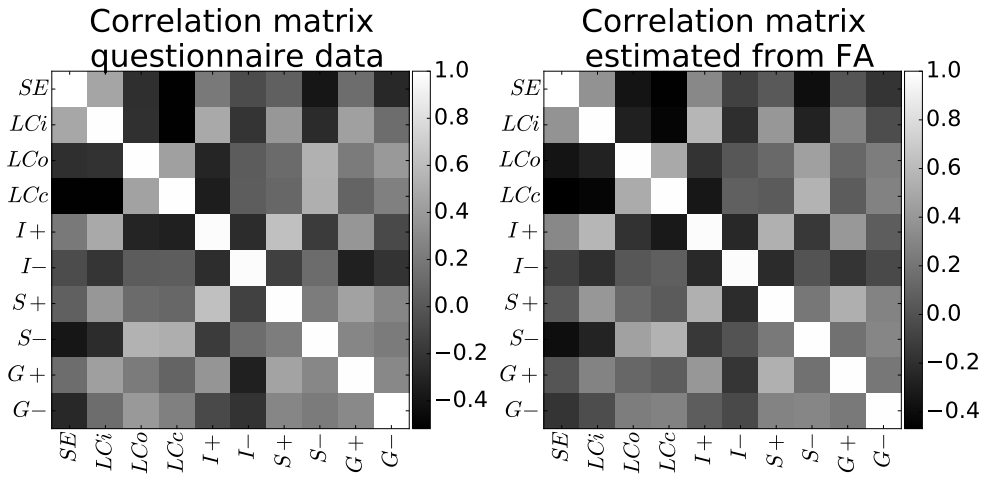


Figure 5.4: Correlation matrix, all questionnaire measures. Left: data correlation matrix. Right: correlation matrix estimate from factor analysis.

5.1.4 Dimensionality reduction: factor analyses

The full set of questionnaire measures was a set of 10-dimensional observations. Figure 5.4 shows the correlation matrix of these data.

We performed factor analysis for dimensionality reduction. Comparison of models with one, two and three latent factors favoured the two factor model: the first eigenvalue of the covariance matrix of the full questionnaire data accounted for 58.77% of the variance, and the hypothesis of one factor was rejected by a chi-squared test($\chi(35) = 67.53, p = 7 * 10^{-4}$); the first three eigenvalues accounted for 91.3% of the variance, but three factors produced overfitting, as indicated by vanishing specific variances; in contrast, the first two eigenvalues accounted for 81.74% of the variance, and the chi-squared test failed to reject the null hypothesis that the total number of factors was two ($\chi(26) = 28.38, p = 0.34$) (see figure 5.4 for comparison between the real correlation matrix and the one estimated from factor analysis).

Table 5.1 shows the resulting factor loadings [3]. Factor 1 (F1) has large negative loadings from self esteem and internal control and large positive loadings from powerful others, chance locus of control and the negative sta-

[3]Factors were rotated with the varimax rotation, an orthogonal rotation which tends to produce sparse, and therefore more interpretable, factor loadings

Factor loadings		
Questionnaire dimension	Factor 1 (F1)	Factor 2 (F2)
SE	-0.55	0.22
LCi	-0.44	0.62
LCo	0.64	-0.01
LCc	0.8	-0.15
I+	-0.28	0.75
I-	0.05	-0.29
S+	0.23	0.82
S-	0.72	0.07
G+	0.2	0.61
G-	0.39	0.24

Table 5.1: Factor loadings, questionnaire scores.

bility subscale; factor 2 (F2) has large positive loadings from internal control and positive internality, stability and globality. We interpret F1 as capturing a dimension of negativity and lack of control, and F2 as capturing a dimension of internal control and generalised positivity.

5.2 Questionnaire scores and behaviour

There were a number of aspects of subjects' behaviour that we hypothesised would be correlated with their questionnaire scores.

Specifically, we expected subjects with higher self-esteem scores to provide higher estimates of skill, to display larger differences in skill updates after wins vs after losses, and to display larger differences between the proportions of internal attributions for wins vs losses, compared to subjects with lower self-esteem. We also expected to find a correlation between scores for internal locus of control and the proportion of internal attributions, as well as between scores on the internality positive and internality negative subscales of the ASQ and the proportions of internal attributions post wins and losses respectively. We computed correlations between the relevant behavioural variable and relevant questionnaire score for or each of these hypothesised relationships and tested significance by performing permutation tests.

With only one exception, our predictions were not confirmed, correla-

tions being generally very weak (see figure 5.5). Only one of the correlations was significant and remained so after correcting for multiple comparisons, namely the correlation between the ASQ internality negative score and the proportion of internal attributions for losses. (r^2 = 0.19, permutation test p-value corrected = 0.04)[4].

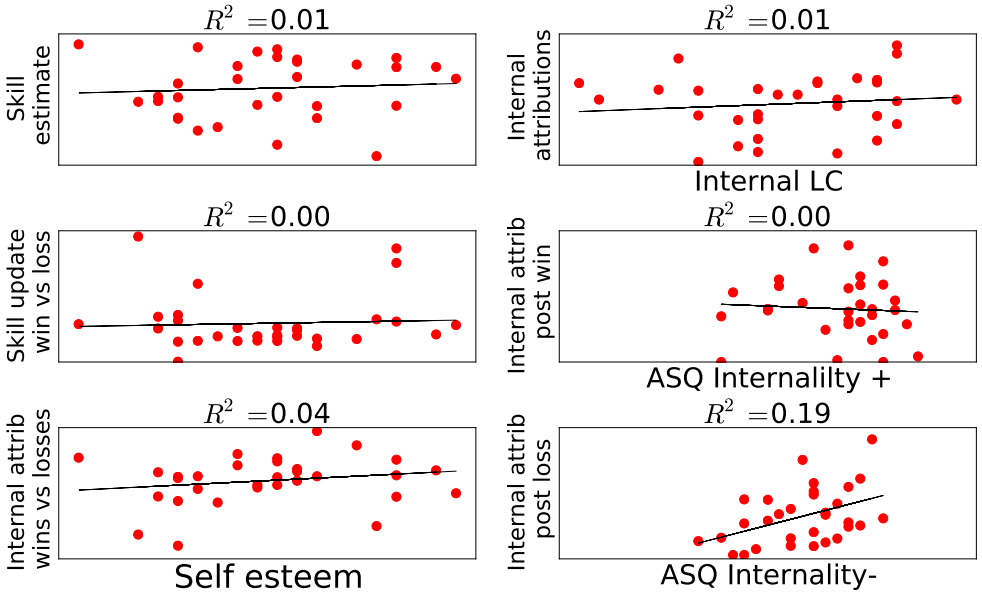


Figure 5.5: Hypothesised correlations between questionnaire scores and behavioural variables of interest, "self" condition. Left hand side: correlations with self esteem scores. Top row: average skill response. Middle row: average difference between skill updates post wins vs losses. Bottom row: difference between the proportion of internal attributions post wins vs losses. Right hand side: Top row: internal locus of control scores and proportion of internal attributions, Middle row: internality positive ASQ scores vs proportion of internal attributions post wins. Bottom row: internality negative ASQ scores vs proportion of internal attributions post losses.

We also performed exploratory analyses[5], investigating correlations between these behavioural measures and the questionnaire scores' projections on the latent factors identified through factor analysis. We found two variables with significant negative correlations with F1, which captures a dimension of negativity and lack of control: the proportion of internal attributions

[4]We also computed correlations with corresponding behavioural measures from the "other" condition, as well as with differences between behavioural measures in the two conditions, with similar results.

[5]We did not perform multiple comparisons corrections for exploratory analyses.

($r^2 = 0.24, p = 0.005$), and the proportion of internal attributions for wins ($r^2 = 0.13, p = 0.047$). We repeated these analyses for behavioural measures from the "other" condition, and found significant negative correlations between F1 and the proportion of internal attributions ($r^2 = 0.22, p = 0.01$), as well as between F1 and the proportion of internal attributions for losses ($r^2 = 0.17, p = 0.02$).

Finally we performed canonical correlation analysis between the behavioural measures from both conditions and the full set of questionnaire scores, in order to determine whether any relationships exist between linear combinations of the behavioural measures and linear combinations of questionnaire measures. We found no significant relationship.

5.3 Questionnaire scores and model parameters

In the remainder of this chapter we present analyses of relationships between model parameters and questionnaire responses. Model-dependent analyses of skill estimates did not provide a clear winning model (see 3.3), therefore we did not perform any further parameter analyses. However this was not the case for models of attribution responses, where we found a clear winning model (see 4.3). In the following we present our analyses investigating relationships between questionnaire measures and individual parameters of the best attribution model.

We were a priori interested in relationships between questionnaire responses and the effect of reported skill on internal attributions in the "self" condition. We begin by presenting the result of these analyses and discuss our interpretation of them, caveats, and the merits of alternative explanations.

We then present results of exploratory analyses of the full set of correlations between parameters and questionnaire measures. These do not correspond to a priori questions, but are aligned with one of two directions. The first involves comparisons that provide evidence for consistency between response behaviour in our task and questionnaire-based measures. The second

type of correlations point to further directions of study into the mechanisms involved in attribution making.

We performed permutation tests to compute approximate p-values for all tested correlations; we performed Bonferroni correction (Bonferroni, 1936) for the a priori hypotheses, and all results reported as significant are reported with Bonferroni-corrected p-values; we did not, however, correct for multiple comparisons in the exploratory analyses, for which significance was established using a threshold of 0.05. We report Pearson correlation coefficients, but we repeated permutation tests for Spearman correlations, with similar results.

5.3.1 Relationships between skill effect and questionnaire measures

In chapter 4 we presented evidence that subjects' reported beliefs about skill contributed to explaining their causal attributions for the outcomes they experienced. One of the purposes of modelling attribution data was to investigate relationships between the effect of skill on subsequent attributions in the self condition [6] and aspects of attribution-making measured by questionnaires.

For most subjects[7], skill had a positive effect of internal attributions for wins and a negative one on internal attributions for losses(see figure 5.6). As increasing skill increases the likelihood of taking credit for wins and decreases the likelihood of taking responsibility for losses, the effect of skill can be

[6] As discussed before (see 4.3.3), individual subject parameters corresponding to raw feature weights do not constitute meaningful measures of the effect of features on attribution responses. This is because features do not exert their effect on attributions directly and independently via the corresponding weights, but through the relationships between these weights and other individual subject parameters (see model description in 4.3). We therefore quantified the subject-level effect of a given feature f on a given attribution option a by computing, on every trial t, the derivative of the probability of response a with respect to f when all remaining features are held constant, $\frac{\partial p_t(a)}{\partial f}$, and averaging over trials (see appendix S for a detailed account). We used this approach to quantify the effect of skill on internal attributions for wins and losses and we computed correlations between these effects and the relevant questionnaire measures, as well as correlations with latent factors.

[7] There were 4 subjects with the opposite effect for losses, and 8 with the opposite effect for wins; due to the small number of subjects, we processed data from all subjects together in the analyses reported below, however investigating differences between subjects with positive and negative skill effects is an avenue for future work.

interpreted as producing attributions that are more favourable to the self.

We were a priori interested in whether these effects are correlated with self-esteem, LCi and the internality subscales of the ASQ questionnaire. We found positive correlations of the effect of skill on internal attributions for losses with LCi ($r = 0.51, p = 0.02$) and I+ ($r = 0.49, p = 0.048$), but none of the other correlations survived corrections for multiple comparisons. Given that skill effects on internal attributions for losses were negative, positive correlations between skill effect and LCi and I+ mean that higher LCi and I+ are associated with a weakening of the effect of skill, as illustrated in figure 5.6.

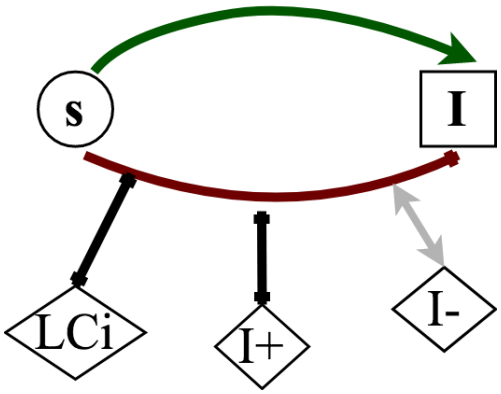


Figure 5.6: Relationships between questionnaire measures and effect of skill on internal attributions. s: skill; I: probability of making internal attributions; LCI, I+, I-: the respective questionnaire measures. Green arrows: effects for wins; red arrows: effects for losses. Pointed arrow heads indicate an 'excitatory' effect (increasing skill increases the likelihood of internal attributions for wins), while blunt arrow heads indicate an 'inhibitory' effect (increasing skill decreases the likelihood of internal attributions for losses). Black and gray arrows: correlations between questionnaire measures and effects; black: significant correlations; gray: correlations that did not survive multiple comparisons corrections. Pointed black and gray arrow heads indicate positive correlations between questionnaire measures and effect strength, blunt arrow heads indicate negative correlation between questionnaire measure and effect strength.

We note that both these correlations seem to be driven to a large extent by three subjects (see figure 5.7). These do not appear to be clear outliers, however we cannot rule out, with the present dataset, the possibility that the connections between LCi and I+ scores and the effect of skill are spurious. Larger datasets are needed to establish whether the relationship is indeed

present, and to determine whether it is a continuous one, or it is due to the existence of clusters of subjects with associated extreme values of LCi and I+ and slope effects.

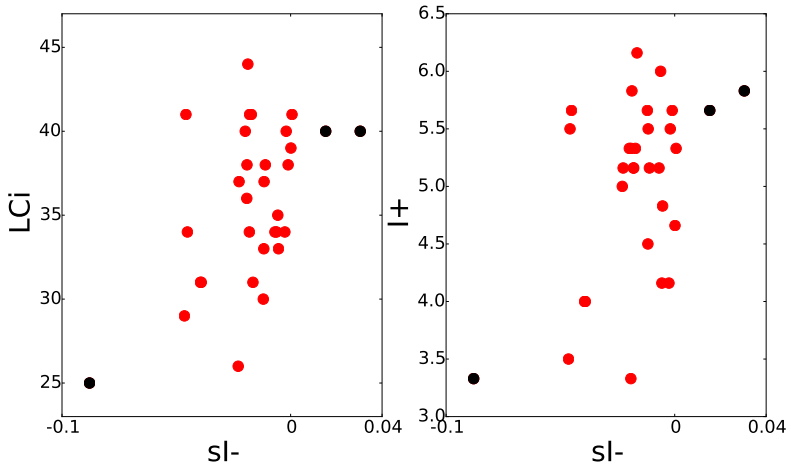


Figure 5.7: Scatter plot of the effect of skill on making internal attribution to losses (sI-) vs the LCi and I+ scores, 'self' condition. Black dots represent the subjects driving most of the correlations. Left: LCi vs sI- Pearson r = 0.51, p = 0.02 for all subjects, r = 0.26, p = 0.11 excluding the three marked subjects. Right: I+ vs sI- r = 0.49, p = 0.048 for all subjects, r = 0.24, p = 0.21 excluding the three marked subjects.

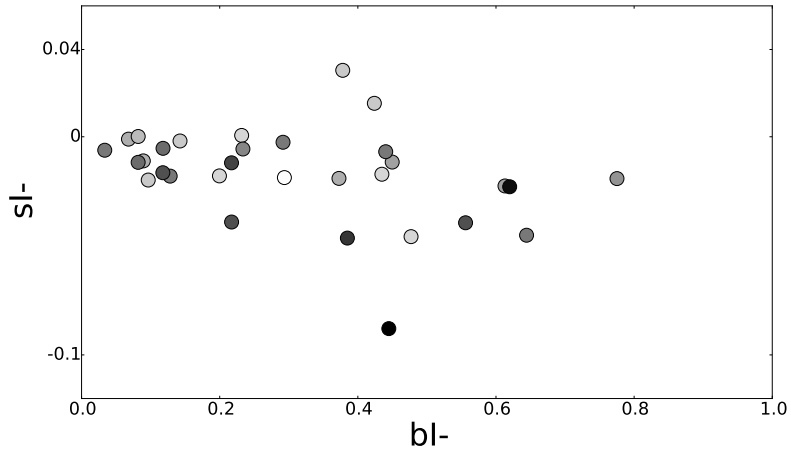


Figure 5.8: The effect of skill for internal attributions for losses (sI-) vs the probability of making internal attributions for losses as computed from the bias term in the model (bI-). Each dot represents a subject, with subjects coloured according to their LCi scores.

We also considered whether the association between LCi and the skill effect might be a mere floor effect, driven by a negative correlation between

LCi and the tendency to make internal attributions for losses: in this scenario, subjects with high LCi scores would tend to make few internal attributions for losses in general and therefore there would be little room for skill to reduce this tendency even further. This does not seem to be the case, however, see figure 5.8. The situation for the relationship with I+ is similar.

The association between LCi and skill effect could be interpreted in terms of responsibility: the higher the score for internal control, the less subjects used skill to decline responsibility for losses. However in this case, assuming that the I- subscale is related to the extent to which subjects perceive internal control for negative outcomes, we would expect the relationship between skill effect and I- scores to mirror that between skill effect and LCi. This is not what we observed (see 5.6): if anything, I- displayed the opposite correlation with skill effect ($r = -0.35$, uncorrected $p = 0.02$), although it did not survive corrections for multiple comparisons.

This pattern of results does not support a coherent interpretation of the data, and further work, as well as a larger population of subjects, are needed to understand the relationships between the two types of measures, and the underlying phenomena. For further discussion, including discussion of alternative mechanisms suggested by the relationships we observed in the data, see section 5.3.2.2.

5.3.2 Exploratory analyses

Relationships between questionnaire measures and the effect of skill on internal attributions were of particular interest to us. However, the way other features (of the task or of performance) contribute to subjects' causal attributions could also potentially be related to the dimensions measured by the questionnaires we administered. We therefore performed exploratory analyses, investigating all possible correlations between the full set of questionnaire measures (and their latent factors) and the effect of all features on all attribution options. We performed permutation tests to estimate the significance of each of the 432 resulting correlations. Since these are exploratory analyses,

we did not perform corrections for multiple comparisons.

We present in this section a selection of the relationships we identified as significant, grouped into two categories: one includes correlations between questionnaire measures and bias parameters capturing propensities for the different attribution options, which show consistency between questionnaire responses and responses in our task; the second includes correlations between questionnaire measures and effects of features other than skill, which suggest potential mechanisms involved in attribution-making. We speculate on the interpretation of these relationships and discuss evidence for and against potential explanations that further work could set out to test.

5.3.2.1 Consistency with questionnaire measures

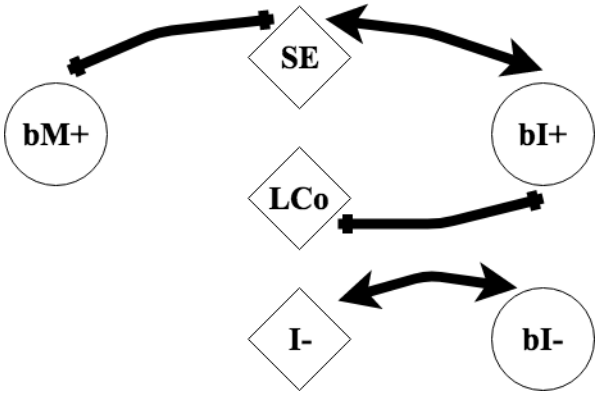


Figure 5.9: Representation of relationships showing consistency between bias parameters in our model and questionnaire measures. bI-, bI+, bM+: bias for making internal attributions for losses, internal attributions for wins and attributions to maze for wins respectively. SE, LCo, I-: the respective questionnaire measures. Pointed arrow heads indicate positive correlations (the higher the SE score is, the higher the propensity for making internal attributions for wins is): blunt arrow heads indicate negative correlations (the higher the LCo score is, the lower the propensity for making internal attributions for wins is).

We found that higher SE scores were associated with a stronger propensity toward making internal attributions for wins -as captured by the bias parameter- ($r = 0.33$, $p = 0.03$) and a weaker propensity towards attributing wins to the maze ($r = -0.36$, $p = 0.02$). Conversely, higher LCo scores were associated with weaker tendency toward attributing wins internally ($r = -0.33$,

p = 0.04), as were higher scores on the latent factor F1 (the "external negative" factor) (r = -0.33, p = 0.04). Finally, higher I- scores were associated with increased probability of attributing losses internally (r = 0.46, $p < 0.01$). See figure 5.9 for an illustration of these relationships.

5.3.2.2 Candidate modulating mechanisms for attributions

In this section we present correlations between questionnaire measures and feature effects that exploratory analyses identified as significant, and which suggest potential mechanisms contributing to subjects' attribution-making. These are broadly grouped according to three main factors: responsibility, maintaining positive beliefs about the self, the role of beliefs about the world.

Figure 5.10 shows the features effects involved in the following analyses, namely effects of skill and performance features on attributions for losses and effects of objective task features on attributions for wins. These effects were consistent with our expectations, for both wins and losses. We briefly review these effects before discussing their relationships with questionnaire measures. Feature effects involved

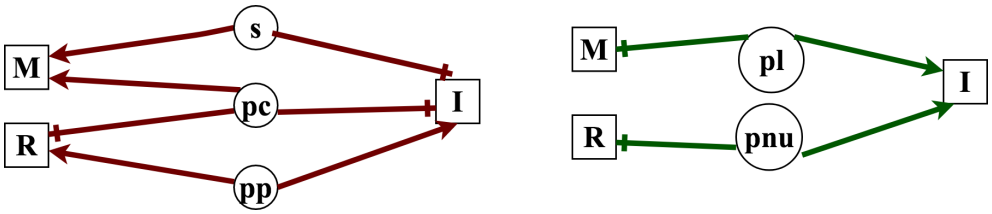


Figure 5.10: Features effects on attributions for losses. Left: attributions for losses. Right: attributions for wins. I: internal attributions, M: attributions to maze, R: attributions to rotations, s: skill, pc: proportion of correct key presses, pp: proportion of pauses, pl: path length, pnu: proportion of non-up orientations. Pointed arrow heads represent positive effects, blunted arrow heads represent negative effects.

As far as attributions for losses are concerned, increasing skill or the proportion of correct key presses (pc) decreases the likelihood of internal attributions, and increases the likelihood of blaming the maze for these outcomes. In addition, the pc feature decreases the likelihood of attributing losses to rotations; since the only way rotations could affect performance is through

making subjects press wrong keys, it is reasonable that losses associated with high proportions of correct presses should not be blamed on rotations. The proportion of pauses (pp) has positive effects on internal attributions and on attributions to rotations for losses, consistent with the fact that higher proportions of pauses can be seen to indicate both poor performance and more disruption caused by rotations.

As far as attributions for wins are concerned, the effects involved are the effect of path length (pl) and proportion of non-up orientations (pnu) on internal attributions and attributions to maze and rotations respectively. Both features had positive effects on internal attributions - the more difficult the task, the more likely subjects were to take credit for wins. Both features had negative effects on the corresponding attribution option: higher difficulty on each direction was associated with lower likelihood of crediting the respective task aspect for wins.

Responsibility

The feature effects on attributions for losses displayed in figure 5.10 can be interpreted as using performance features to divert blame for losses away from self and towards external options. Subjects' patterns of taking responsibility for outcomes might be related to some of the correlations that we observed between these effects and questionnaire measures. We would expect responsibility to be associated with a weakening of effects related to avoidance or deflection of blame, and a strengthening of effects related to assuming blame internally (see figure 5.11).

We mentioned above the correlation between LCi and the effect of skill on internal attribution for losses as a potential marker for the role of responsibility in attribution making: higher LCi scores were associated with decreased use of skill to avoid blame for losses. Exploratory analyses revealed other correlations between feature effects and questionnaire measures, consistent with this view: both LCi and I- were negatively correlated with the extent of using performance to avoid internal blame for losses (see figure 5.12, left and right). There were also significant correlations with the latent factor F2 (which has

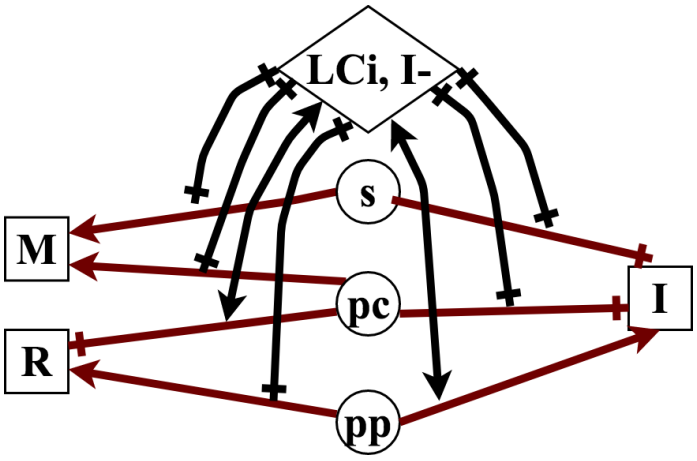


Figure 5.11: Representation of relationships that would reflect a responsibility mechanism driving internal attributions. See figure 5.12 for the relationships we observed in our data.

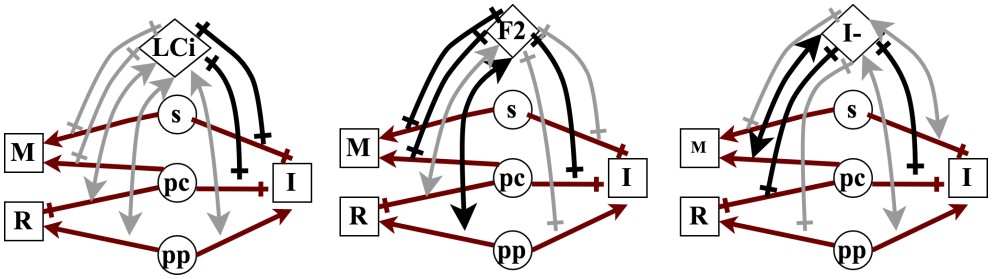


Figure 5.12: Representation of observed relationships involving variables related to responsibility. I: internal attributions, M: attributions to maze, R: attributions to rotations, s: skill, pc: proportion of correct key presses, pp: proportion of pauses. Pointed arrow heads represent positive effects, blunted arrow heads represent negative effects. Black arrows represent correlations between these effects and questionnaire measures. Left: Higher LCi dampens the effect of skill (previously mentioned r = 0.51, p = 0.02) and pc ($r = 0.37$, $p = 0.02$) on internal attributions; none of the other correlations with LCi is significant. Middle: We found several significant correlations between the latent factor F2 (the 'internal positive' factor, which has a large LCi loading) and these effects. Thus F2 dampens the effect of skill on attributions to maze (r= -0.33, p = 0.04) and the effects of the pc feature on both internal attributions ($r = 0.4$, $p = 0.01$) and attributions to maze ($r = -0.32$, $p = 0.04$). Finally F2 boosts the effect of the pc feature on attributions to rotations ($r = -0.34$, $p = 0.03$). None of the other relationships involving F2 is significant. Right: I- dampens the effects of the pc feature on both internal attributions ($r = 0.4$, $p = 0.01$) and attributions to rotations ($r = -0.34$, $p = 0.03$) and boosts the effect of pc on attributions to maze ($r = -0.4$, $p = 0.01$). None of the other correlations with I- is significant. We note that the thresholds for significance for relationships with the effects of skill on internal attributions are more stringent than the other thresholds, as they are Bonferroni corrected for multiple comparisons.

large loading from LCi) consistent with this view: negative correlations between F2 and the extent of using pc to avoid attributing losses internally, as well as negative correlations between F2 and the extent of using skill and pc to blame the maze for losses (see figure 5.12 middle).

However, we would expect responsibility to be associated with other correlation patterns that we did not see in our data. See figure 5.11 for an illustration of the relationships expected to arise from responsibility. First, if I-can be interpreted as measuring subjects perception of their internal control in situations producing negative outcomes, we would expect it to mirror LCi in its correlations with the effects of the features of interest, which was not the case. Secondly, we would expect responsibility to be correlated with the effects of the pp feature on attributions, which was also not the case. Finally, F2 has large loadings from questionnaire measures not related to responsibility, therefore further work is needed to disentangle the contributions that other factors have in its relationships with the effects of various features.

Positive belief maintenance

Some of the correlations we identified point to another set of mechanisms which are likely to be involved in attribution making: processes maintaining subjects' positive beliefs about the self. From this perspective, there are two types of patterns that we might expect. One includes positive correlations between markers of vulnerability such as low SE or high I- and taking credit for wins or avoiding blame for losses; such associations might be interpreted as revealing an increased need for the activation of processes that can maintain a positive self-image. Alternatively, the second type involves positive correlations between SE scores and patterns of taking credit for wins or avoiding blames for losses; these might be interpreted as revealing the contribution of causal attributions toward successful maintenance of positive beliefs.

We found evidence consistent with both of these phenomena. The correlations between I- and the effect of skill and pc on internal attributions for losses (see figure 5.12, right) are consistent with the first one, according to which higher vulnerability of positive beliefs is associated with a heightened

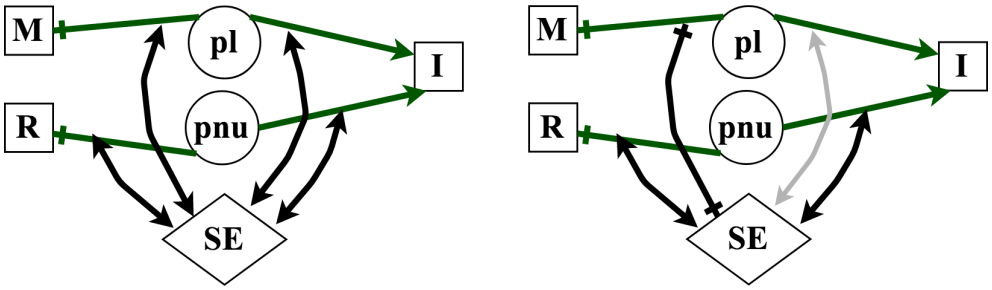


Figure 5.13: Representation of relationships that would reflect a mechanism for maintaining positive beliefs about the self. Left: expected relationships. Right: relationships observed in our data; gray arrows indicate correlations that were not significant.

need to protect the self by deflecting blame: the higher the I- score, the more subjects used skill and pc to deflect blame from themselves for losses.

We also found relationships consistent with the converse view, according to which attributions supporting positive beliefs are associated with higher self-esteem. One pattern of attribution which likely contributes to positive self-image involves the effect of the proportion of unusual orientations (pnu feature) on internal attributions and attributions to rotations: increasing the proportion of unusual orientations increases the tendency to attribute wins internally and decreases the tendency to attribute them to rotations. We found significant positive correlations of SE with the strength of both these effects (effect on internal attributions: $r = 0.37$, $p = 0.02$, effect on attributions to rotations $r = -0.47$, $p < 0.01$[8]); see figure 5.13. However we might expect a similar relationship to hold for SE and the effect of the pl feature on internal attributions and attributions to the maze, and this was not the case in our data. See figure 5.13 for an illustration.

Finally, some of the correlations detected by exploratory analyses (see figure 5.14) suggest a role of beliefs that subjects hold about the world in modulating the way information about performance and task features is processed for making causal attributions.

Beliefs about the world

[8]Note that the effect of pnu on attributions to rotations is negative, therefore a negative value of the correlation with SE indicates that higher SE scores are associated with stronger effects.

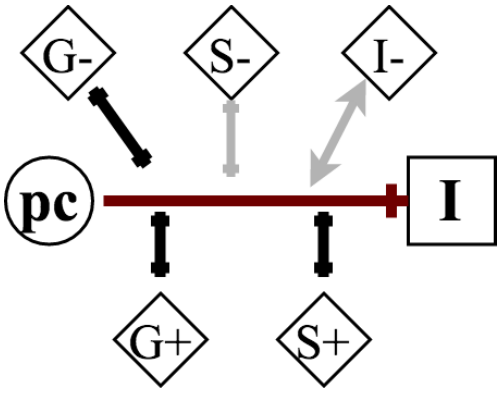


Figure 5.14: Representation of relationships suggesting a role of beliefs about the world in attribution-making. Gray arrows indicate correlations that were not significant.

We found that the extent to which the pc feature decreases internal attribution for losses was negatively correlated with scores on both globality subscales (G+ :$r = 0.41$, $p = 0.01$, G-: $r = 0.53$, $p < 0.01$), as well as with the S+ subscale ($r = 0.35$, $p = 0.02$[9]). The correlations with G+ and S+ could be interpreted as a reflection of the effect that holding positive views about the world has on causal attributions: an optimistic belief that the world is generally favourable and stable could make blaming external causes for failures less likely.

Alternatively, holding pessimistic views about the world could be associated with a depressive failure of the protective effect of pc, in which case we would expect negative correlations between the strength of its effect on internal attribution for losses and G-, S- and potentially I-. As mentioned above, we found a significant correlation with G-, however correlations with S- ($r = 0.16$, $p = 0.2$) and I- ($r = -0.28$, $p = 0.06$) did not support this explanation.

We have presented in this section a number of correlations between ques-

[9]Note that the effect of pc on internal attributions for losses is negative, therefore a positive value of the correlation with questionnaire measures indicates that higher scores are associated with weaker effects.

tionnaire measures and the effects of various features on attribution responses. They provide inconsistent evidence, whose reliability and indeed validity is difficult to establish given the limited amount of data available. We did not have a priori hypotheses about these relationships, and we cannot, with the present data, disentangle between the diverse potential factors contributing to causal attribution making.

Rather, we conceived of these exploratory analysis as useful in suggesting hypotheses that future analysis could set out to investigate in a targeted way. The relationships that we observed indicate three such putative mechanisms as promising candidates for future research: responsibility, processes maintaining positive beliefs about the self, subjects' beliefs about the world. Selective manipulation - of subjects' control over the outcomes, the strength of their belief in their own abilities, and the volatility of the task - is needed to disentangle the contributions of the three putative mechanisms to causal attributions.

5.3.2.3 'Self' vs 'other'

A priori, we were also interested in whether any putative relationship between questionnaire responses and model parameters in the "self" condition would be present in the "other" condition, or whether there is any systematic change in such relationships between the two conditions. Therefore for all correlations that exploratory or a priori analyses identified as significant in the "self" condition, we tested their correspondents in the "other" condition.

We found that all but two of these relationships were not significant for "other", the two exceptions being a negative correlation between SE and the effect of pnu on attributing wins to rotations ($r = -0.52,\ p < 0.01$) and a negative correlation between F2 and the effect of pc on attributing losses to rotations ($r = -0.3,\ p = 0.05$), both consistent with their counterparts for "self".

We note that there are two major difficulties in interpreting relationships between questionnaire responses and parameters obtained by fitting data in

the "other" condition.

The first is that the questionnaires that we used refer to attitudes and beliefs held by the self, and centered around the self; if others are referenced (as is the case in the "others" subscale of the Levenson's questionnaire), they are referenced with respect to the self, which is unlike the situation in our task, in which subjects are asked to evaluate the "other" as an independent agent. This is why we did not have precise a priori expectations of particular relationships between questionnaire responses and parameters related to attribution responses in the "other" condition.

The second difficulty is that we do not have a precise measure of the extent to which subjects suspected our deceit about the "other". Further studies of such differences are needed to establish to what extent they are indeed present in the population, and to better understand them.

5.4 Discussion

There were three main reasons for administering questionnaires to our subjects.

First we sought to determine whether responses in our task show meaningful expected correlations with the respective questionnaires. Questionnaires involve hypothetical situations relating to self evaluation and causal attributions, whereas our task involves making causal attributions for real experienced events, and reporting beliefs which should be forming and changing during the task. Therefore relationships between task responses and questionnaire responses constitute both a test of the extent to which questionnaires can be used to predict real responses in such a context, and, to the extent that these questionnaires are accepted as tapping into meaningful and relevant psychological dimensions of attribution making and belief updating, a test of the extent to which our task allows access to some of the same phenomena.

Evidence about the expected relationships between descriptive statistics of the data and questionnaire measures was mixed: we could identify some,

but not all, as significant; however further model parameter analyses provided additional evidence in favour of such links. As is the case with all results obtained from these data, additional investigation on larger data sets are necessary to fully establish the validity of these observations; we believe there is enough evidence to suggests that such additional investigations are warranted.

Our second aim was to investigate whether attribution making, and in particular the effect that belief about skill has on it, is modulated by any of the dimensions measured by questionnaires. We found some evidence that this is the case in the "self" condition, as indicated by significant correlations between questionnaire measures (LCi and I+) and the effect of skill on internal attributions. As noted before, this observation comes with the caveat that these relationships seem to be mostly driven by a few subjects, and that, with the present data, we cannot establish to what extent they are a mere artefact or reveal real underlying mechanisms, perhaps driven by clusters of subjects. This is one of the questions that additional data would be able to address.

Finally, in line with our conception of this experiment as a pilot for future investigations, we aimed to identify potential mechanisms and formulate more precise research questions that future research could directly address. Exploratory analyses of relationships between model parameters in the "self" condition and questionnaire measures indicated several potential mechanisms that could contribute to or modulate attribution making.

Subjects' patterns of internalising responsibility might be involved in the processing of skill and performance information, modulating the extent to which these features are used to blame external factors, instead of the self, for bad outcomes.

Alternatively or additionally, blame avoidance might be related to the need for preserving a set of positive beliefs about the self. This could either be as a healthy mechanism, in which case it would be associated with high scores on scales indicating positive beliefs about the self, or as an indication of an increased need for preserving vulnerable self esteem, in which case it

would be associated with other measures of vulnerability.

Our observations also hinted at the importance of more general beliefs about the world: it might be more difficult to blame the external world for one's failures, if one holds the belief that the world is mostly benevolent. Conversely, generally negative beliefs about the world are known to be associated with a depressive mood, this being indeed one of the theories that the ASQ was designed to test.

We found mixed evidence about some of these factors from our exploratory analyses of correlations between model parameters and questionnaire measures. Here, too, additional data would hopefully provide a clearer picture, and manipulations could be used to disentangle the contribution of different aspects. Manipulating the world volatility, for instance, would affect the contribution that beliefs about the world have on blame assignment, while having a lesser effect , if any, on the responsibility factor. Manipulating subjects' beliefs about themselves by exposing them to artificial experiences of failure or success could be used to specifically affect the contribution that the need for preservation of self esteem has on driving causal attributions. And manipulating the extent of control that subjects have over their performance could be used to specifically impact the contribution of the responsibility factor to attribution making.

We conclude that the current data obtained from this task shows promising consistency with existing validated questionnaire measures, and indicates interesting avenues for future, more precisely targeted work.

Chapter 6

Conclusions and future work

In this chapter we present a summary of the thesis, highlighting the original contributions of this work; we then present our perspectives on future work, from task improvements that our analyses suggested to a broader outlook on further reated research directions.

6.1 Summary and original contributions

The work presented in this thesis was inspired by the theoretical proposal put forth by Bentall et al (Bentall et al., 2001; Bentall, 2003)- the attribution-self-representation cycle theory, and rooted in the computational approach to learning and decision making. Our aim was to provide a quantitative framework for phrasing, exploring and testing hypotheses associated with this theory. As such, this work is part of a larger recent effort to bring the powerful approach and tools of computational neuroscience to bear on complex, high level aspects of cognition and emotion, and their disruptions in psychiatric disorders. This is to build on the successes of these methods in exploring and explaining various fundamental aspects of decision making in animals and humans (see review in chapter1).

We developed a novel task, in which subjects' beliefs and causal attributions are repeatedly probed with regards to real outcomes, experienced in the context of learning a skill task. This framework allows the collection of time series of attributions and beliefs, which are fundamentally necessary for

investigating, on a trial-by-trial basis, the dynamic relationships postulated by the theory.

We found evidence of effects consistent with the cycle postulated by the theory, namely effects of attributions on beliefs about skill and effects of beliefs about skill on attributions, neither of which could be reduced to the effect of objective performance. Crucially, in our work these effects were observed at trial-level resolution, providing evidence for moment-by-moment interactions that could support the sort of complex dynamics postulated by the theory (Bentall et al., 2001; Bentall, 2003). In addition, the beliefs and causal attributions involved were reported by subjects in the absence of any manipulation aimed at biasing or otherwise controlling them[1].

The attribution-self-representation cycle theory is fundamentally a theory about the self, however the reasoning mechanisms that it postulates could also be involved in the formation and maintenance of beliefs about others. Our task provides a framework within which the dynamics of beliefs and attributions about the self can be directly compared with their counterparts regarding another. We found evidence of the same interactions between beliefs about skill and attributions in both conditions, although data from from the "self" condition was generally noisier and some aspects of the task appeared to be processed differently between the two conditions (see below).

Finally, we found novel evidence of consistency between patterns of casual attributions provided for real outcomes experienced in the task and questionnaire-based measures of control and attributional style, which rely on appraisal of hypothetical situations.

Thus our work proposed and implemented a novel framework for investigations into the dynamics of beliefs and causal attributions, and completed the first steps towards a precise formalisation and testing of the theoretical proposal in this framework. Analyses of our data showed that naturally oc-

[1] Subjects were deceived about the identity of the "other", in order to allow controlled comparison with the "self" condition (see 2 for details), but no manipulation was aimed at the content of their causal attributions, or their beliefs about skill, in either condition.

curring beliefs and causal attributions interact on a moment-by-moment basis, both when the self is concerned, and when appraising others, thus providing novel evidence in support of the theoretical proposal's validity.

6.2 Reflections on the task

Since our task is novel, there were a number design choices that we could not readily base on evidence from previous work. Some of these inadvertently introduced requirements for external validation, which we could only approximate to a limited extent with our available data; others introduced confounds or undesirable complications for data analysis. We reflect on these with the benefit of hindsight, highlighting improvements that could be implemented in future work. We discuss more general directions for future work in the subsequent section.

To begin with, there were no previously validated objective measures of difficulty and skill in our task. However adapting trial difficulty to subjects' performance level was necessary in order to maintain the balance between positive and negative outcomes, as well as to maintain subjects' engagement in the task. For this purpose we designed a rather complicated staircase procedure, involving the various aspects of the task that were, intuitively, the most relevant, and then extracted an objective measure of difficulty from the data. We also attempted to extract an objective measure of skill from our data, but not being satisfied with the quality of the resulting measure, we decided not to use it further (see 2). The staircase procedure did achieve the practical desiderata of satisfying time constraints and providing balanced numbers of wins and losses, however it failed to closely track subjects' performance level, producing instead undesirably high trial-to-trial variations in difficulty(see 2). In this work we focused on establishing that the overall framework is viable, and did not pursue precise staircase calibration or the external validation of an objective difficulty measure; these remain goals for future work. They can both be achieved by systematically exploring the space of task variables

acting as components of difficulty, and measuring the frequency of wins and losses for each setting of these variables in a large population of subjects.

Using a precisely calibrated staircase in conjunction with an externally validated measure of difficulty would allow skill to be objectively measured in a relatively straightforward way, since in this case difficulty (as manipulated by the staircase) could be assumed to closely track subjects' true skill levels. This would be a useful task improvement, allowing comparisons to be made between subjects' skill estimates and their real underlying skill. The accuracy of subjects' beliefs about skill could then be investigated in relationship with their patterns of attributing outcomes, or questionnaire-based psychological measures. A number of questions that we could not address in this work could then be asked, such as whether subjects display self-serving biases in their skill estimates, and if that is the case, whether such biases are associated with self-serving attribution patterns, or with higher self-esteem scores. Indeed we observed large between-subjects variability in subjects' reported skill estimates, and model-dependent analyses indicated that variability persisted in mechanisms underlying skill updates; having access to an objective measure of skill, along with larger datasets, would enable further investigations into the nature and sources of this variability.

We probed causal attributions by multiple choice questions. The response options for these questions were designed to allow investigation of internal vs external attributions, as well as attributions to different aspects of the task that we could objectively measure, such as maze complexity and rotation frequency. However, the specific response options we provided introduced an availability bias - there were three external, but only one internal options, with the latter being unitary and the former being more fine-grained. In addition, we provided response options for maze complexity and rotations as independent candidate causes, however in hindsight this might not have matched subjects' perception. Subjects experiencing the task might have perceived the corresponding task aspects as being of different natures, allowing

different degrees of control and being hierarchically organised, rather than independent. Indeed both model agnostic and model-based analyses indicated that there were differences in the way they were processed. Future task iterations could balance internal and external attribution options by providing similarly fine-grained options for both, highlighting specific aspects of one's skill (or lack thereof) that mirror specific task difficulties: e.g. 'my skill in dealing with rotations', 'my lack of skill in following the correct path' etc. Furthermore, subjects could be asked to report the degree to which they attribute the outcome to each of the available factors, rather than having to pick one of them.

Truthful responding to the skill and attribution questions was not incentivised. Our attempt to use an incentive-compatible measurement -BDM betting (Becker et al., 1964)- alongside the direct question about skill turned out to be unsuccessful (see 2). In order to avoid subjects' losing interest and becoming demotivated to respond truthfully to skill and attribution questions, as well as due to time constraints, we only asked subjects for skill estimates and attributions every second trial. We therefore did not have complete trial-by-trial series of both attributions and skill estimates, which hindered some of our analyses (see 3). Finer calibration of the staircase procedure resulting in more precise tracking of subjects' skill level would contribute to solving this issue, by allowing trials to more efficiently explore the relevant regime. Reducing trial numbers could allow us to solicit skill and attribution assessments on every trial.

Finally, we introduced the "other" condition to allow direct comparison of skill estimates and attributions for self with those provided for another, while controlling for difficulty and trial order. We implemented this condition by recording subjects' trials and playing them back, with minimal modifications aimed at reducing the likelihood of their being recalled. This resulted in a fixed "self" -"other" order which we could not control for, and in a need to deceive our subjects. This latter aspect complicates interpretations of dif-

ferences between the two conditions, as it is difficult to precisely establish the extent to which subjects suspected deceit, and the timing of their suspicions, if any. Future work is needed for thorough investigations of the psychophysics of the task; this understanding, along with better calibration of difficulty, could conceivably enable artificial simulation of agents with specified properties, removing present constraints on the "other" condition, and enabling additional questions to be tackled.

6.3 Perspectives on future work

The work presented in this thesis proposes a framework in which the attributional-self representation cycle theory (Bentall et al., 2001; Bentall, 2003) can be quantitatively defined, refined and tested. The theory postulates the existence of two-way interactions between beliefs about the self and appraisal of events, and highlights the dynamic nature of these variables, which, if coupled, can give rise to complex dynamics. According to the theory, the system is calibrated to maintain adaptive beliefs and causal attribution patterns, enabling humans to deal with everyday adversities without significant damage to their self esteem, self concept or mental health. It is, however, also able to support aberrant dynamics: a combination of temporary vulnerability and severe adversity can push the system out of its normal regime into vicious circles, resulting in depression or paranoia.

As the name clearly states, the core concept of the theory is the cycle linking the two variables. In this work we have investigated the two sides of the cycle independently. We added to the existing evidence on the dynamic nature of both beliefs and attributions (Forgas et al., 1990; Bentall and Kaney, 2005; Dunning et al., 1995), and we have provided evidence of one-way effects of each variable on the other, at the trial-level time resolution. These are necessary first steps, laying the ground for further research aimed at understanding the cycle dynamics and their involvement in mental health and disorders.

One direction for future research, indeed perhaps the natural next step, involves investigating the range of behaviours that the coupling between attribution and beliefs about self can support, and the extent to which different underlying mechanisms can be distinguished from observable patterns (Eldar and Niv, 2015).

As the example in chapter 1 showed, coupling between variables can amplify randomly occurring variations in the task and the agent's behaviour, producing qualitatively different behavioural patterns, even in very simplified situations: in that case, random differences in attributions at the beginning of the task were amplified by the 2-way connections between attributions and beliefs, leading to different preferences between the two available actions. The complexity available in our task presumably allows for a broader spectrum of behaviours: subjects can experience both wins and losses, and there are multiple causes they could attribute these outcomes to. Future work could explore this space by simulating the interactions between the two variables while varying parameters such as sensitivity to positive and negative outcomes, levels of noise in making internal and external attributions, levels of noise in estimating skill, and the strength of the couplings between attributions and beliefs. Simulations could thus provide a better understanding of the observable effects that various parameters can produce, and the nature of theoretical predictions that can be tested within this context.

Analyses presented in this work were performed under the assumption that the functional effects of the two variables on each other were constant in time, as well as independent of the variables' levels. Relaxing this assumption is another goal for future work. Here as well, simulations and modelling work could be used to explore the possible dynamics supported by interactions which can themselves evolve in time, or depend on the values of the interacting variables. This extension would allow the phrasing and testing of hypotheses involving multiple time-scales of interaction, as well as hypotheses involving threshold-like mechanisms that might be responsible for catas-

trophic dynamics. As an example, one can imagine testing, in this context, whether an association between particularly low levels of beliefs about self and particularly strong effects of negative attributions on beliefs could model the onset of depression (Bentall, 2003) .

This direction is naturally related to another goal for future research, which is particularly relevant given the inspiration for this work: understanding the effects of disruptions or interventions on the system. Simulation analyses could be further expanded to investigate the nature of perturbations and or environment manipulations that can impact agents' beliefs and causal attributions. Thus artificial experiments could be performed to study a number of aspects of interest, such as the effects of repeated small losses vs isolated exceptionally large losses, the effects of losses concentrated vs distributed in time (and their analogues for wins), whether providing alternative attributions for particularly significant events can change the dynamics etc. Factors pushing the system towards catastrophic vicious circles are of particular interest, as are protective factors that prevent or dampen such dynamics (Robins and Hayes, 1995; Liu et al., 2015; Haeffel and Vargas, 2011; Johnson et al., 2017).

Insights from such simulated experiments could be used to optimise experiment design and test hypotheses in real populations.

The directions discussed above are directly suggested by the theory itself. The pattern of results we observed in our data also indicated interesting avenues for future work.

We repeatedly observed differences between the processing of wins and losses (Seymour et al., 2007; Frank et al., 2007; Cools et al., 2008; Sharot et al., 2011), both in the updating of beliefs about skill, and in the effects of attributions. These observations suggested that accurate assignment of blame for losses might be privileged w.r.t. accurate assignment of credit for wins - a plausible phenomenon from an evolutionary viewpoint, as the cost of wrong inferences about negative outcomes can be significantly higher than that incurred for inaccurate inferences about gains. Future work could test

the merit of this hypothesis by manipulating the environment, in particular its level of benevolence/danger, and the relative importance of accurate causal attributions for positive and negative outcomes. Manipulations of the environment itself could include pre-exposure to hostile vs benevolent environments (Iigaya et al., 2016), or rigging the task itself in favour or against subjects (Forgas et al., 1990; Bentall and Kaney, 2005). The relevance of making correct inferences could be modified by allowing subjects to ask for changes in particular aspects of the environment based on their causal attributions, or to choose between alternative trials varying on the direction indicated by their attributions; by making the effects of these choices on actual task changes probabilistic, with varying levels of reliability for wins and losses, the relative importance of accuracy for the two outcomes could be controlled.

In our task, we found differences between the effects of path length and rotations on attributions to the relevant options, as well as differences between these effects in the "self" vs "other" conditions (see 4). One candidate explanation for these differences involves the degree of perceived control over these task aspects, and the extent to which subjects estimate control differently when playing vs when watching. Control is intimately connected with the core notion of attribution, and another avenue for future work involves directly manipulating the level of control subjects have in the task (Mancinelli et al., 2020), and investigating the effects of such manipulations on subjects' beliefs, attributions and interactions between them. This could be achieved by introducing noise in the execution of subjects' commands, as well as by allowing subjects to directly control speed, e.g. by use of joystick rather than key pressing. As far as differences between "self" and "other" are concerned, introduction of multiple "other" conditions could be used to control for the effect of watching vs playing.

We also observed systematic differences between data in the "self" and "other" conditions, generally consistent with a noisier pattern of skill responses for self, alongside a pattern of more favourable judgement of the

other. As mentioned above, task improvements could remove ordering confounds and the need for deceit, allowing further investigation to establish whether these effects are indeed present. There are a number of additional directions that future work could explore with regards to the "self"-"other" distinction. One such direction involves the effect of the "otherness" of the other on any differences with respect to self: thus evaluation of close others vs indifferent, hostile or artificial others could be compared. Further expansion into the social aspect of the processes involved can also be envisioned, by introducing competition into the task, or by providing agents with a learning environment populated by real or artificial peers.

Finally, exploratory analyses of the relationships between questionnaire measures and mechanisms inferred from model-dependent analyses pointed to three directions of interest: subjects' patterns of internalising responsibility, pressures to maintain positive beliefs about the self and the importance of beliefs about the world. Correlational analyses on larger datasets and experimental manipulations could be used to investigate these phenomena and test alternative hypotheses about them. Responsibility mechanisms could be probed by manipulating control, while pressure to maintain positive beliefs could be altered via artificial success or failure experiences prior to the task. The role of beliefs about the world could be investigated by manipulating aspects of the task environment, such as its benevolence/malevolence (also mentioned above in connection with learning from wins vs losses) and its volatility.

We hope the work presented in this thesis proves useful to those investigating these and other related questions, and provides motivation and inspiration for future research.

Appendix A

Instructions condition self

In this experiment, you will learn how to navigate a maze under unusual conditions. The experiment involves two related tasks, which you'll perform around a week apart. Each of the tasks is about 1.5-2 hours long, so it will be divided in two parts, which you can do on the same day, with a break of a few hours, or on two successive days.

In the first task, which you're starting now, you will navigate the maze and make decisions based on your own performance. In the second task, which will take place later, you will watch and make decisions based on someone else navigating the maze. We will tell you about that then.

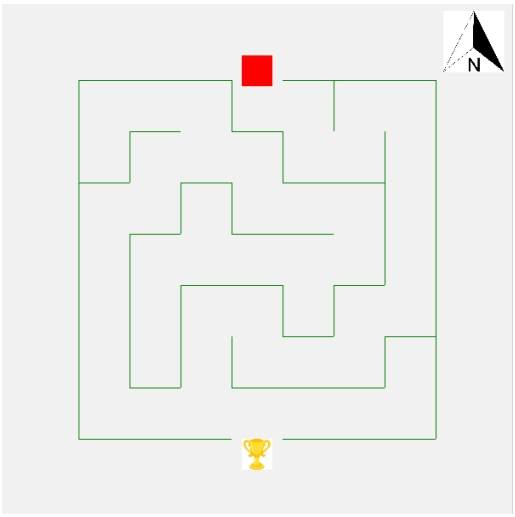
THE TASK

On each trial, you have to move a red square through a maze towards a goal marked by a trophy. Each trial uses a different maze, some of which are more difficult than others; the difficulty is randomly chosen on every trial. You win if you manage to get to the goal in the available time, otherwise you lose.

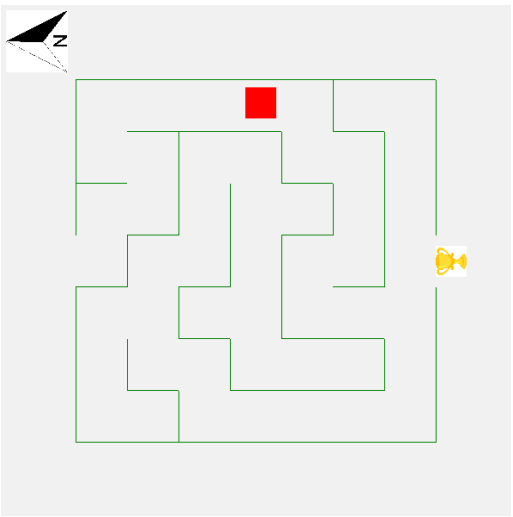
At the end of the session, 5 out of all the trials will be picked randomly. Your trial payment will be £1 for each trial that you won out of these 5. You will be offered the opportunity to gain additional bonus reward, as detailed below.

You move the red square by using the arrow keys on the keyboard. Each key moves the red square in a cardinal direction: the UP key moves the red square to the North, the RIGHT key moves it to the East, the DOWN keys moves it to the South, the LEFT key moves it to the West. There is a compass needle in the corner of the screen, marking North. Every once in a while during a trial, the screen will rotate. When this happens, the compass needle will rotate together with the maze. Pressing the arrow keys will still move the red square towards the cardinal directions indicated by the compass, but now North might no longer be UP on the screen.
Here are two screen shots, one at the beginning of the trial (A) and one later in the trial, after a rotation (B).

A



B



In screen shot B, pressing UP moves the red square North, which is to the left, pressing DOWN moves it to the right, pressing RIGHT moves it up and pressing LEFT moves it down.

You will practise 10 trials before the start of the experiment, so that you can become familiar with the task and the rules of the navigation problem. Whether you win or lose on these trials won't count towards your final payment.

PART 1: YOU PLAY

Once you finish the practice trials, the experimental trials start. From this moment on, any of the trials could be among the 5 trials chosen randomly at the end, which determine your trial payment.

Before one out of every 3 trials, you will be given the opportunity of winning an additional bonus reward whose maximum value (£1) is equivalent to winning one trial.

Normally, this bonus would depend on your success on the forthcoming trial (so £1 or £0 if you win or lose, respectively). You can think of this opportunity of wining the bonus as a gamble. We would like to know how much you value the possibility of winning this bonus, and so you will engage in an auction. In the auction, you will specify the smallest amount between £0 and £1 that you would accept as a sure amount in exchange for this gamble. We call this your 'minimum price'.

The computer then draws randomly a value between £0 and £1.

If the value is larger than your minimum price, then you 'sell' the gamble, and you get the value the computer drew as a bonus, irrespective of whether you win or lose the trial.

If the value the computer draws is smaller than your minimum price, then you don't 'sell' the gamble, and then whether or not you gain the bonus will depend on whether you win or lose the trial.

In this kind of auction, the best thing for you is to tell us the true value that the gamble has for you. Here's why:

Suppose the gamble is really worth 70p to you.

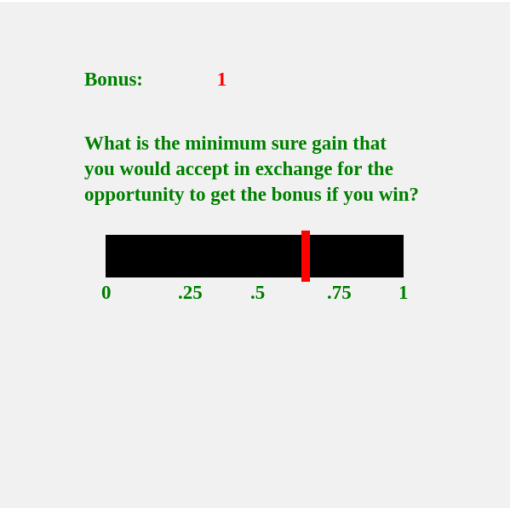
If you say 50p, then if the computer draws 60p, you trade the gamble for the computer value, and you end up with 60p, which is less that the value of the gamble itself.

If, on the other hand, you tell us that your minimum price is 90p, then if the computer's value is 80p, you don't sell the gamble. So you end up with the gamble, which is worth 70p, instead of winning the computer drawn 80p.

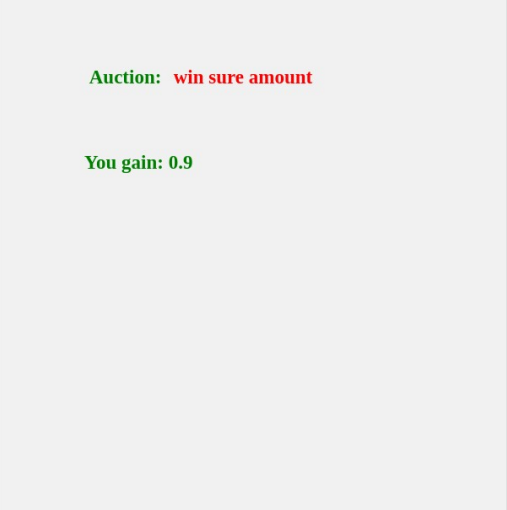
If you tell us your real value, then there is no way for you to end up with something that is worth less for you than the gamble, and you also don't miss any opportunity.

Here is an example of how this auction will look in the task. Suppose you say the minimum price that you would accept is 70p. (C) After you play the trial, you get to see the result of the auction. The computer generates 90p, which is larger than your minimum price. Then you have traded the bonus for a sure amount, and you get the price drawn by the computer, which is 90p. (D)

C



D



You will also be asked, from time to time, about how difficult or easy the task seems to you.
You will enter all your responses by using the arrow keys. Use the left and right arrow keys to move the slider, then press Enter to save and move on.

Appendix B

Instructions condition other

This is the second task in the experiment. It is very similar to the task you performed already. The only difference is that now instead of navigating the maze yourself, you will watch someone else do it. Now it is the performance of this other person that determines your gains, and so you will have to make decisions based on their performance.

THE TASK

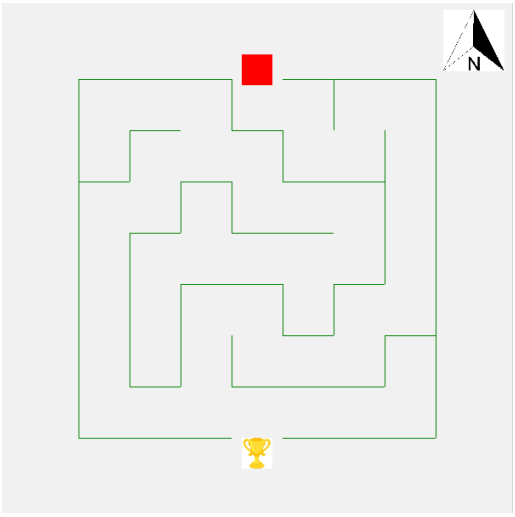
On each trial, the subject whose performance you are watching, let's call him or her X, has to move a red square through a maze towards a goal marked by a trophy. Each trial uses a different maze, some of which are more difficult than others; the difficulty is randomly chosen on every trial. X wins if he/she manages to get to the goal in the available time, otherwise he/she loses.

At the end of the session, 5 out of all the trials will be picked randomly. **Your** trial payment will be £1 for each trial that X won out of these 5. You will be offered the opportunity to gain additional bonus reward, as detailed below.

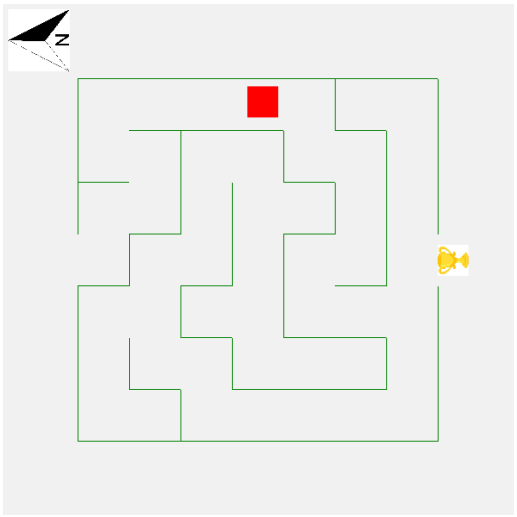
Just as you did in the first task, X moves the red square by using the arrow keys on the keyboard. Each key moves the red square in a cardinal direction: the UP key moves the red square to the North, the RIGHT key moves it to the East, the DOWN keys moves it to the South, the LEFT key moves it to the West. There is a compass needle in the corner of the screen, marking North. Every once in a while during a trial, the screen will rotate. When this happens, the compass needle will rotate together with the maze. Pressing the arrow keys will still move the red square towards the cardinal directions indicated by the compass, but now North might no longer be UP on the screen.

Here are two screen shots, one at the beginning of the trial (A) and one later in the trial, after a rotation (B).

A



B



In screen shot B, pressing UP moves the red square North, which is to the left, pressing DOWN moves it to the right, pressing RIGHT moves it up and pressing LEFT moves it down.

PART 2: X PLAYS

Before one out of every 4 trials, you will be given the opportunity of winning an additional bonus reward whose maximum value (£1) is equivalent to winning one trial.

Normally, this bonus would depend on X's success on the forthcoming trial (so £1 if X wins, £0 if X loses). You can think of this as a gamble. We would like to know how much you value the possibility of winning this bonus, and so you will engage in an auction. In the auction, you will specify the smallest amount between £0 and £1 that you would accept as a sure amount in exchange for this gamble. We call this your 'minimum price'.

The computer then draws randomly a value between £0 and £1.

If the value is larger than your minimum price, then you 'sell' the gamble, and you get the value the computer drew as a bonus, irrespective of whether X wins or loses the trial.

If the value the computer draws is smaller than your minimum price, then you don't 'sell' the gamble, and then whether or not you gain the bonus will depend on whether X wins or loses the trial.

In this kind of auction, the best thing for you is to tell us the true value that the gamble has for you. Here's why:

Suppose the gamble is really worth 70p to you.

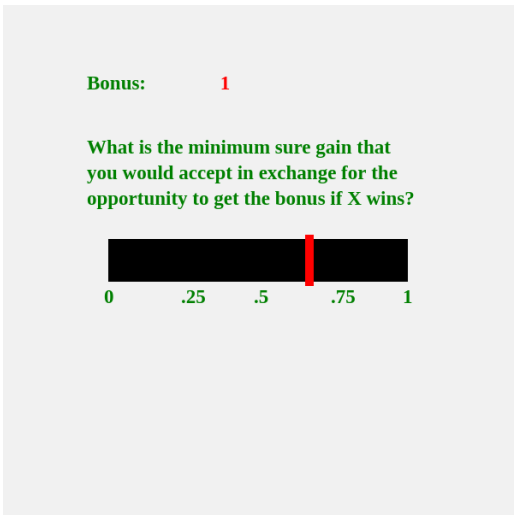
If you say 50p, then if the computer draws 60p, you trade the gamble for the computer value, and you end up with 60p, which is less that the value of the gamble itself.

If, on the other hand, you tell us that your minimum price is 90p, then if the computer's value is 80p, you don't sell the gamble. So you end up with the gamble, which is worth 70p, instead of winning the computer drawn 80p.

If you tell us your real value, then there is no way for you to end up with something that is worth less for you than the gamble, and you also don't miss any opportunity.

Here is an example of how this auction will look in the task. Suppose you say the minimum price that you would accept is 70p. (C) After X plays the trial, you get to see the result of the auction. The computer generates 90p, which is larger than your minimum price. Then you have traded the bonus for a sure amount, and you get the price drawn by the computer, which is 90p. (D)

C



D

You will also be asked, from time to time, about how difficult or easy you think the trials are for X and how difficult or easy you would find them if you played yourself.

You will enter all your responses by using the arrow keys. Use the left and right arrow keys to move the slider, then press Enter to save and move on.

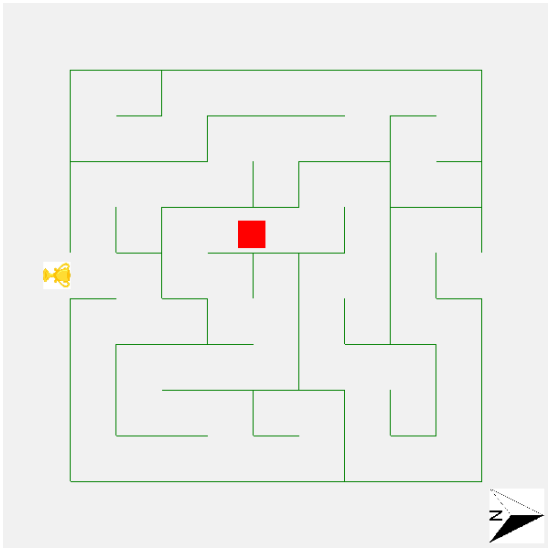
Appendix C

Verification questions

QUESTIONS

1. In the following situation, where does the red square move if you press the DOWN arrow key?

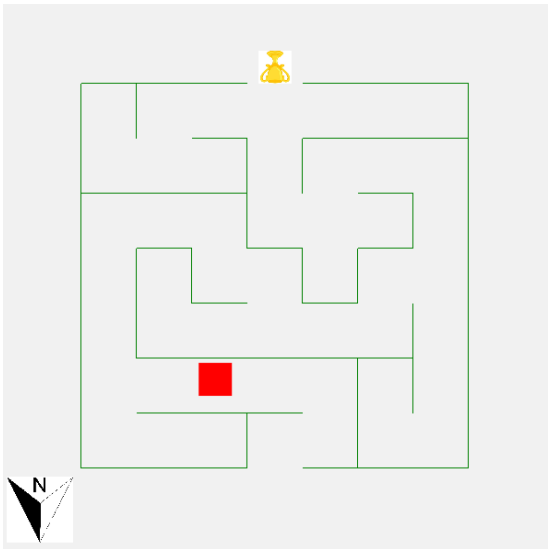
What happens if you press the RIGHT arrow key?



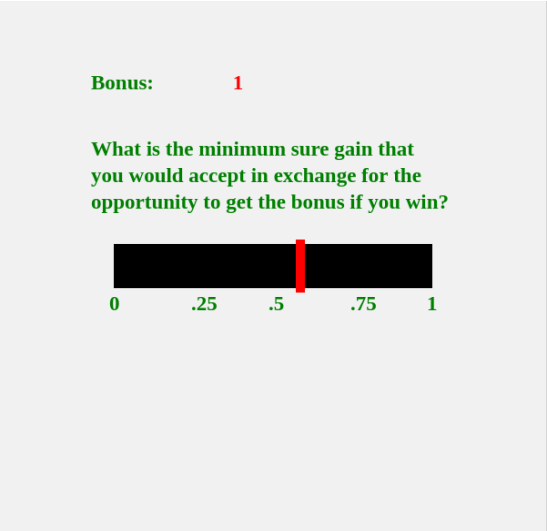
2. In the following situation, what key would you press to move the red square to the right?

Draw an arrow on the maze to indicate the correct first movement towards the goal.

What key would you press to move the red square towards the goal?



3. Suppose you choose 0.6 in the following situation and you win the trial.
If the computer generated random number is 0.8 do you win any bonus reward? If yes, how much?



4. What if you choose 0.4 and you lose the trial, but the computer generated number is 0.7?
Do you win any bonus then? If yes, how much?

What if the computer-generated number is 0.3?

Appendix D

Feedback questions

Debriefing questionnaire

For each of the following statements, indicate the extent to which you agree or disagree by ticking the appropriate box.

	Strongly disagree -3	**Disagree** -2	**Slightly disagree** -1	**Slightly agree** +1	**Agree** +2	**Strongly agree** +3
1. I found watching more engaging than playing.						
2.I did not use any strategy for betting on X's performance.						
3. I feel my performance and X's performance were similar.						
4. I felt I was evaluating myself more harshly than I was evaluating X.						
5. I feel I became better and better at the task while playing it.						
6. I felt I could predict quite accurately if I would win or lose the next trial.						
7. I used a strategy for betting on my performance.						
8.I found the task engaging.						
9. I feel I had easier trials than X.						
10. I feel X was better than me at this task.						
11. I found the task boring.						
12. I feel I was better than X at this task.						
13. I feel X didn't make much progress throughout the task.						
14. I felt I was evaluating X more harshly than myself.						
15. I often felt there was nothing I could do to win when I was playing.						

16. I felt I could predict quite accurately if X would win or lose the next trial.						
17. I found playing more engaging than watching.						
18. I found the task frustrating.						
19. I felt I was evaluating myself in a fair way.						
20. I felt I was evaluating both X and myself fairly.						
21. I felt the task was getting harder and harder for me.						
22. I often felt I could win if I was focused enough when I was playing.						
22. I felt the task was getting easier and easier for me.						
23. I feel I had a more difficult job than subject X.						
24. I felt I learned faster than X.						

Briefly describe your strategy for betting on your trials:

Briefly describe your strategy for betting in X's trials.

Mark your overall estimates of your own and X's skills at this task:

---

very bad very good

Do you have any other comments? Or questions?

Some of our participants have thought that some of subject X's trials were actually played-back versions of their own trials. What percentage of subject X's trials did you think might have been your own?

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Appendix E

Questionnaires

ID: _____

Questionnaire 1

For each of the following statements, indicate the extent to which you agree or disagree by ticking the appropriate box.

	Strongly disagree -3	**Disagree** -2	**Slightly disagree** -1	**Slightly agree** +1	**Agree** +2	**Strongly agree** +3
1. Whether or not I get to be a leader depends mostly on my ability.						
2. To a great extent my life is controlled by accidental happenings.						
3. I feel like what happens in my life is mostly determined by powerful people.						
4. Whether or not I get into a car accident depends mostly on how good a driver I am.						
5. When I make plans, I am almost certain to make them work.						
6. Often there is no chance of protecting my personal interests from bad luck.						
7. When I get what I want, it's usually because I'm lucky.						
8. Although I might have good ability, I will not be given leadership responsibility without appealing to those in positions of power.						
9. How many friends I have depends on how nice a person I am.						
10. I have often found that what is going to happen will happen.						
11. My life is chiefly controlled by powerful others.						

	Strongly disagree -3	**Disagree** -2	**Slightly disagree** -1	**Slightly agree** +1	**Agree** +2	**Strongly agree** +3
12. Whether or not I get into a car accident is mostly a matter of luck.						
13. People like myself have very little chance of protecting our personal interests when they conflict with those of strong pressure groups.						
14. It's not always wise for me to plan too far ahead because many things turn out to be a matter of good or bad fortune.						
15. Getting what I want requires pleasing those people above me.						
16. Whether or not I get to be a leader depends on whether I'm lucky enough to be in the right place at the right time.						
17. If important people were to decide they didn't like me, I probably wouldn't make many friends.						
18. I can pretty much determine what will happen in my life.						
19. I am usually able to protect my personal interests.						
20. Whether or not I get into a car accident depends mostly on the other driver.						
21. When I get what I want, it's usually because I worked hard for it.						
22. In order to have my plans work, I make sure that they fit in with the desires of people who have power over me.						
23. My life is determined by my own actions.						
24. It's chiefly a matter of fate whether or not I have a few friends or many friends.						

ID: _____

Questionnaire 2

Please try to vividly imagine yourself in the situations that follow.

If such a situation happened to you, what would you feel would have caused it? While events may have many causes, we want you to pick only one - the **major** cause if this event happened to **you**. Please write this cause in the blank provided after each event.

Next we want you to answer some questions about the **cause** and a final question about the **situation**.

To summarize, we want you to:

1. Read each situation and vividly imagine it happening to you.
2. Decide what you feel would be the **major** cause of the situation if it happened to you.
3. Write one cause in the blank provided.
4. Answer three questions about the **cause**.
5. Answer one question about the **situation**.
6. Go on to the next situation.

Situation 1: **You meet a friend who compliments you on your appearance.**

1. Write down **one** major cause:

_____

2. Is the cause of your friend's compliment due to something about you or to something about other people or circumstances? (circle one number)

Totally due to other people or circumstances 1 2 3 4 5 6 7 Totally due to me

3. In the future when meeting your friend, will this cause again be present? (circle one number)

Will never again be present 1 2 3 4 5 6 7 Will always be present

4.Is the cause something that just influences this friend opinion on your appearance or does it also

influence other areas of your life? (circle one number)

Influences just this particular situation	1	2	3	4	5	6	7	Influences all situations in my life

5.How important would this situation be if it happened to you?(circle one number)

Not at all important	1	2	3	4	5	6	7	Extremely important

Situation 2: **You have been looking for a job unsuccessfully for some time.**

1. Write down **one** major cause:

_____

2. Is the cause of your unsuccessful job search due to something about you or to something about other people or circumstances? (circle one number)

Totally due to other people or circumstances	1	2	3	4	5	6	7	Totally due to me

3. In the future when looking for a job, will this cause again be present? (circle one number)

Will never again be present	1	2	3	4	5	6	7	Will always be present

4.Is the cause something that just influences looking for a job or does it also influence other areas of your life? (circle one number)

Influences just this particular situation	1	2	3	4	5	6	7	Influences all situations in my life

5.How important would this situation be if it happened to you?(circle one number)

Not at all important	1	2	3	4	5	6	7	Extremely important

Situation 3: **You become very rich.**

1. Write down **one** major cause:

_____

2. Is the cause of your richness due to something about you or to something about other people or circumstances? (circle one number)

Totally due to other people or circumstances	1 2 3 4 5 6 7	Totally due to me

3. In the future, will this cause again be present? (circle one number)

Will never again be present	1 2 3 4 5 6 7	Will always be present

4.Is the cause something that just influences how rich you are or does it also influence other areas of your life? (circle one number)

Influences just this particular situation	1 2 3 4 5 6 7	Influences all situations in my life

5.How important would this situation be if it happened to you?(circle one number)

Not at all important	1 2 3 4 5 6 7	Extremely important

Situation 4: **A friend comes to you with a problem and you don't try to help.**

1. Write down **one** major cause:

_____

2. Is the cause of your not helping due to something about you or to something about other people or circumstances? (circle one number)

Totally due to other people or circumstances 1 2 3 4 5 6 7 Totally due to me

3. In the future when friends come to you with their problems, will this cause again be present? (circle one number)

Will never again be present 1 2 3 4 5 6 7 Will always be present

4.Is the cause something that just influences how you respond to friends' problems or does it also influence other areas of your life? (circle one number)

Influences just this particular situation 1 2 3 4 5 6 7 Influences all situations in my life

5.How important would this situation be if it happened to you?(circle one number)

Not at all important 1 2 3 4 5 6 7 Extremely important

Situation 5: **You give an important talk in front of a group and the audience reacts negatively.**

1. Write down **one** major cause:

_____

2. Is the cause of the audience's reaction due to something about you or to something about other people or circumstances? (circle one number)

Totally due to other people or circumstances 1 2 3 4 5 6 7 Totally due to me

3. In the future when giving talks, will this cause again be present? (circle one number)

Will never again be present 1 2 3 4 5 6 7 Will always be present

4.Is the cause something that just influences your public speaking or does it also influence other areas of your life? (circle one number)

Influences just this particular situation	1 2 3 4 5 6 7	Influences all situations in my life

5.How important would this situation be if it happened to you?(circle one number)

Not at all important	1 2 3 4 5 6 7	Extremely important

Situation 6: **You do a project that is highly praised.**

1. Write down **one** major cause:

_____

2. Is the cause of the praise due to something about you or to something about other people or circumstances? (circle one number)

Totally due to other people or circumstances	1 2 3 4 5 6 7	Totally due to me

3. In the future when working on projects, will this cause again be present? (circle one number)

Will never again be present	1 2 3 4 5 6 7	Will always be present

4.Is the cause something that just influences how your projects are appreciated or does it also influence other areas of your life? (circle one number)

Influences just this particular situation	1 2 3 4 5 6 7	Influences all situations in my life

5.How important would this situation be if it happened to you?(circle one number)

Not at all important	1 2 3 4 5 6 7	Extremely important

Situation 7: **You meet a friend who acts hostilely toward you.**

1. Write down **one** major cause:

_____

2. Is the cause of your friend's hostility due to something about you or to something about other people or circumstances? (circle one number)

Totally due to other people or circumstances	1 2 3 4 5 6 7	Totally due to me

3. In the future when meeting your friend, will this cause again be present? (circle one number)

Will never again be present	1 2 3 4 5 6 7	Will always be present

4.Is the cause something that just influences your interaction with your friend or does it also influence other areas of your life? (circle one number)

Influences just this particular situation	1 2 3 4 5 6 7	Influences all situations in my life

5.How important would this situation be if it happened to you?(circle one number)

Not at all important	1 2 3 4 5 6 7	Extremely important

Situation 8: **You can't get all the work done that others expect of you.**

1. Write down **one** major cause:

_____

2. Is the cause of your not getting all the work done due to something about you or to something about other people or circumstances? (circle one number)

Totally due to other people or circumstances	1 2 3 4 5 6 7	Totally due to me

3. In the future when working, will this cause again be present? (circle one number)

Will never again be present	1 2 3 4 5 6 7	Will always be present

4.Is the cause something that just influences your work or does it also influence other areas of your life? (circle one number)

Influences just this particular situation	1 2 3 4 5 6 7	Influences all situations in my life

5.How important would this situation be if it happened to you?(circle one number)

Not at all important	1 2 3 4 5 6 7	Extremely important

Situation 9: **Your spouse (boyfriend/girl friend) has been treating you more lovingly.**

1. Write down **one** major cause:

_____

2. Is the cause of your spouse treatment of you due to something about you or to something about other people or circumstances? (circle one number)

Totally due to other people or circumstances	1 2 3 4 5 6 7	Totally due to me

3. In the future in your relationship, will this cause again be present? (circle one number)

Will never again be present	1 2 3 4 5 6 7	Will always be present

4.Is the cause something that just influences your relationship or does it also influence other areas of your life? (circle one number)

Influences just this particular situation	1 2 3 4 5 6 7	Influences all situations in my life

5.How important would this situation be if it happened to you?(circle one number)

Not at all important	1 2 3 4 5 6 7	Extremely important

Situation 10: **You apply for a position that you want very badly (e.g., important job, graduate school admission) and you get it.**

1. Write down **one** major cause:

_____

2. Is the cause of your getting the position due to something about you or to something about other people or circumstances? (circle one number)

Totally due to other people or circumstances	1	2	3	4	5	6	7	Totally due to me

3. In the future when applying for positions, will this cause again be present? (circle one number)

Will never again be present	1	2	3	4	5	6	7	Will always be present

4.Is the cause something that just influences your applications or does it also influence other areas of your life? (circle one number)

Influences just this particular situation	1	2	3	4	5	6	7	Influences all situations in my life

5.How important would this situation be if it happened to you?(circle one number)

Not at all important	1	2	3	4	5	6	7	Extremely important

Situation 11: **You go out on a date and it goes badly.**

1. Write down **one** major cause:

_____

2. Is the cause of your date going badly due to something about you or to something about other people or circumstances? (circle one number)

Totally due to other people or circumstances	1	2	3	4	5	6	7	Totally due to me

3. In the future when going on dates, will this cause again be present? (circle one number)

Will never again be	1	2	3	4	5	6	7	Will always be

present present

4.Is the cause something that just influences your dating or does it also influence other areas of your life? (circle one number)

Influences just this particular situation 1 2 3 4 5 6 7 Influences all situations in my life

5.How important would this situation be if it happened to you?(circle one number)

Not at all important 1 2 3 4 5 6 7 Extremely important

Situation 12: **You get a raise.**

1. Write down **one** major cause:

_____

2. Is the cause of your getting a raise due to something about you or to something about other people or circumstances? (circle one number)

Totally due to other people or circumstances 1 2 3 4 5 6 7 Totally due to me

3. In the future, will this cause again be present? (circle one number)

Will never again be present 1 2 3 4 5 6 7 Will always be present

4.Is the cause something that just influences your position at work or does it also influence other areas of your life? (circle one number)

Influences just this particular situation 1 2 3 4 5 6 7 Influences all situations in my life

5.How important would this situation be if it happened to you?(circle one number)

Not at all important 1 2 3 4 5 6 7 Extremely important

ID: _____

Questionnaire 3

Below is a list of statements dealing with your general feelings about yourself. Please indicate how strongly you agree or disagree with each statement.

	Strongly Agree	Agree	Disagree	Strongly disagree
1. On the whole, I am satisfied with myself.				
2. At times I think I am no good at all.				
3. I feel that I have a number of good qualities.				
4. I am able to do things as well as most other people.				
5. I feel I do not have much to be proud of.				
6. I certainly feel useless at times.				
7. I feel that I'm a person of worth, at least on an equal plane with others.				
8. I wish I could have more respect for myself.				
9. All in all, I am inclined to feel that I am a failure.				
10. I take a positive attitude toward myself.				

Appendix F

Staircase procedure

We used a double staircase procedure. Each stair is characterised by three variables:

- the overall size of the maze, n: if the maze is conceived as a square nXn matrix of "maze chambers", with each chamber having four possible walls, which can be present or absent, all the combinations of wall patterns that form a valid maze give the total set of available mazes for a given maze size n; there are four possible "levels" of maze size on our staircases, consisting in two values for n, one of which is randomly drawn, with equal probability, before generating the maze on each trial: level 0 has available n values $\{3,5\}$, level 1 has available n values $\{5,5\}$ level 2 has available n values $\{5,7\}$ and level 3 has available n values $\{7,7\}$.

- the average frequency of maze rotations during a trial $\overline{v}$: all trials start in the normal upright position and the first rotation, resulting in a randomly chosen orientation at an angle of 90, 180 or 270 degrees with respect to the upright one, happens 30 frames (1.5 seconds) later; a random number is then uniformly drawn from the interval $(\overline{v}-10,\overline{v}+10)$, representing the number of frames until the next rotation; the angle of the rotation is drawn randomly with equal probability from the three available options (90, 180, 270 degrees) every time a rotation happens.

The available values for $\overline{v}$ were between 20 and 140 frames, and a staircase step was 10 frames.

- the available time for the trial, t: the available values for t are 10, 12 and 14 seconds, with the staircase step being 2 seconds.

Both staircases started with the same value for the available time, the maximum one of 14 seconds. One of the staircases started on level 1 of the maze dimensions available and on $\overline{v} = 30$ frames, the other started on level 2 of the maze dimensions available and on $\overline{v} = 80$ frames. These values were updated as follows: if still possible, $\overline{v}$ was increased (decreased) by 10 for wins and losses respectively; when the upper (lower) limit was reached, the maze dimension level was increased(decreased), if still possible; when the upper (lower) limit was reached for this as well, the available time t was decreased (increased) if possible.

Appendix G

Difficulty and skill recovery simulations

As discussed in 2.3.1.1, we found that difficulty and skill recovery were successful when simulated staircases used narrow distributions of step sizes, for a large range of placements of these distributions' modes.

In this very simplified, 1-dimensional case, it is only in extreme and unrealistic regimes that recovery breaks down. Exceedingly small step sizes cause difficulty to lag behind skill, which leads to an inversion of the normal relationship between difficulty and the probability of winning(see figure G.1), impairing difficulty (and therefore skill) recovery. Exceedingly large step sizes produce situations in which difficulty alone fully predicts outcomes, and therefore impair skill recovery (see figure G.2 for an illustration of outcome prediction accuracy for different staircase step sizes).

In order to investigate the effect of using highly variable step sizes, we also generated staircase steps from an exponential distribution. We found that in this case too difficulty and skill recovery were successful (see figures G.3 and G.4). We used the exponential distribution for all further simulations.

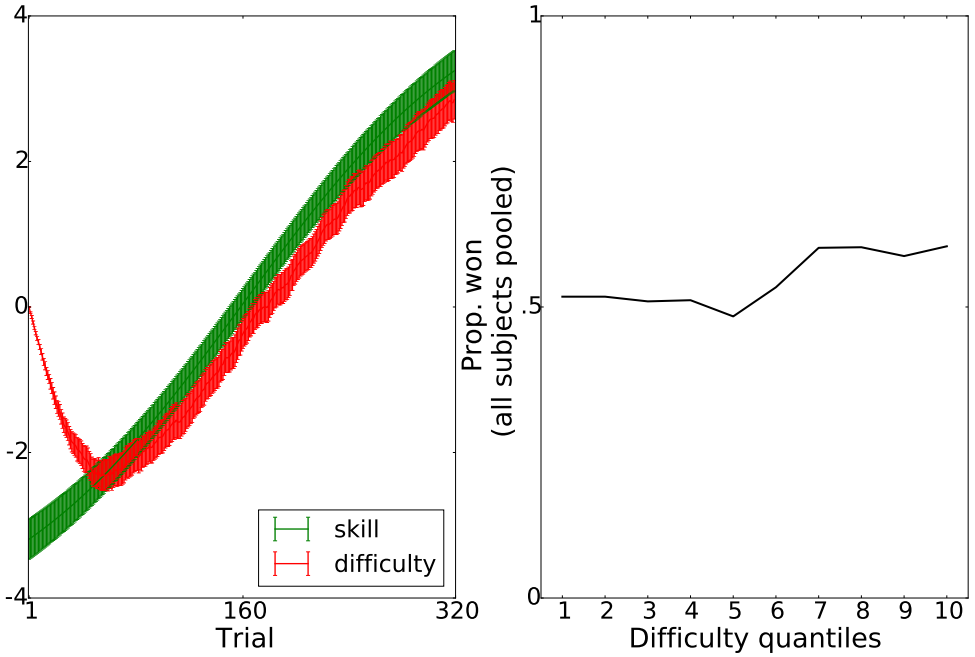


Figure G.1: Effect of exceedingly small staircase steps (parameters used for Gamma distribution: $\theta = 0.005, k = 11$). Left: evolutions of skill and difficulty, mean $\pm$ s.e.m across subjects. Right: relationship between difficulty and outcome distribution, data pooled from all subjects.

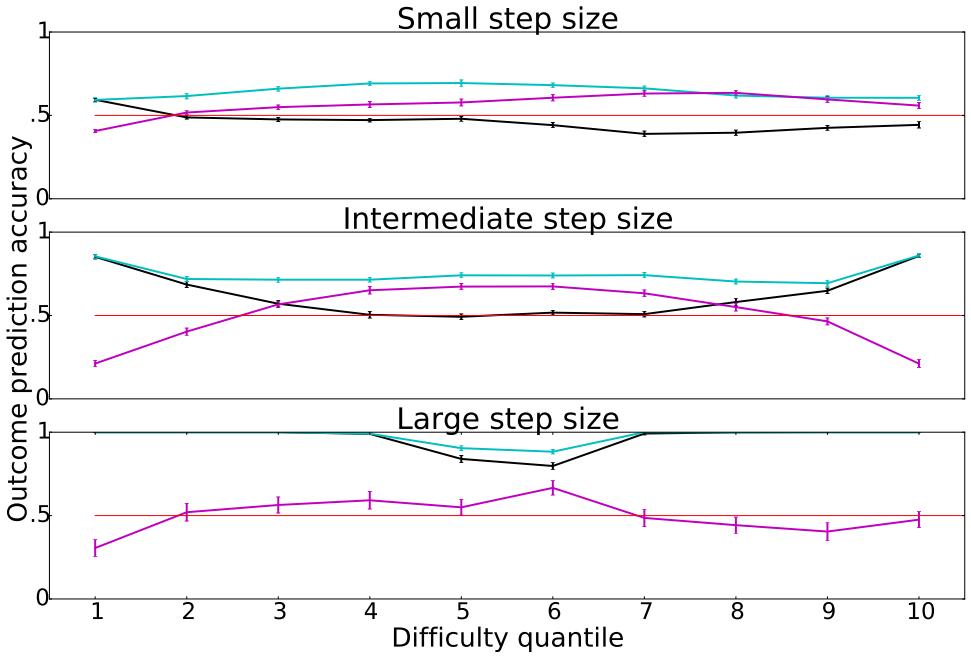


Figure G.2: Outcome prediction accuracy as a function of difficulty level for small (top), intermediate (middle) and large (bottom) average step sizes, mean $\pm$ s.e.m. across simulated subjects. Colours correspond to predicting outcome using difficulty only (black), using skill only (magenta) and using both (cyan).

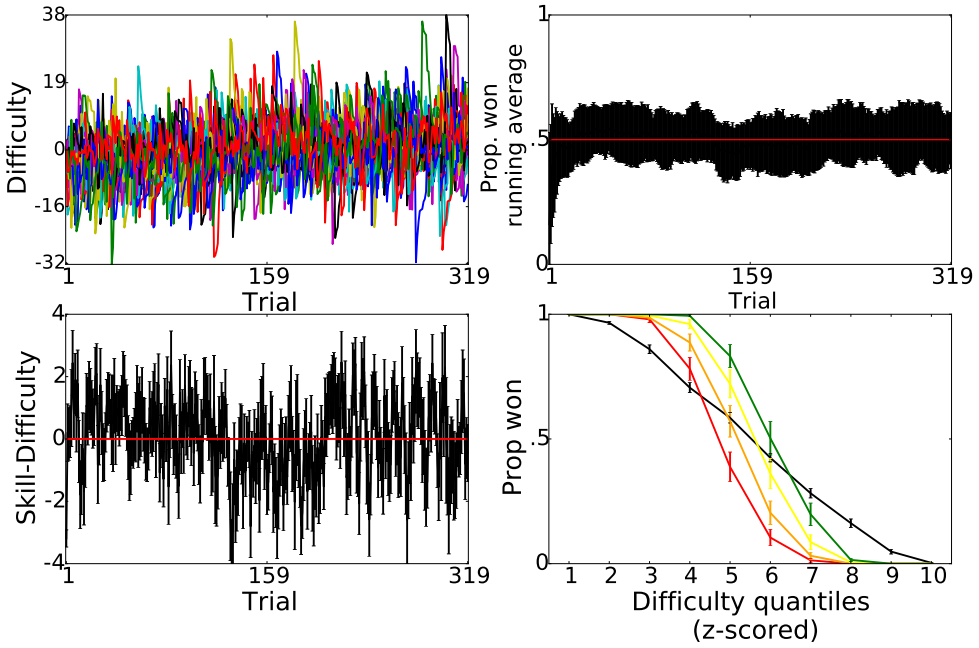


Figure G.3: Example summary of simulated data for step size drawn from exponential distribution (parameter used: $\lambda = 0.4$). Top left: difficulty evolution for all simulated subjects. Top right: running average of the proportion of trials won, mean $\pm$ s.e.m across subjects, filtering window: 20 trials. Bottom left: evolution of difference between skill and difficulty, mean $\pm$ s.e.m across subjects. Bottom right: relationship between difficulty and the proportion of wins, overall and as a function of skill level; mean $\pm$ s.e.m across subjects; black: overall, red to green: 1^{st} to 4^{th} skill quartiles; difficulty z-scored within subject.

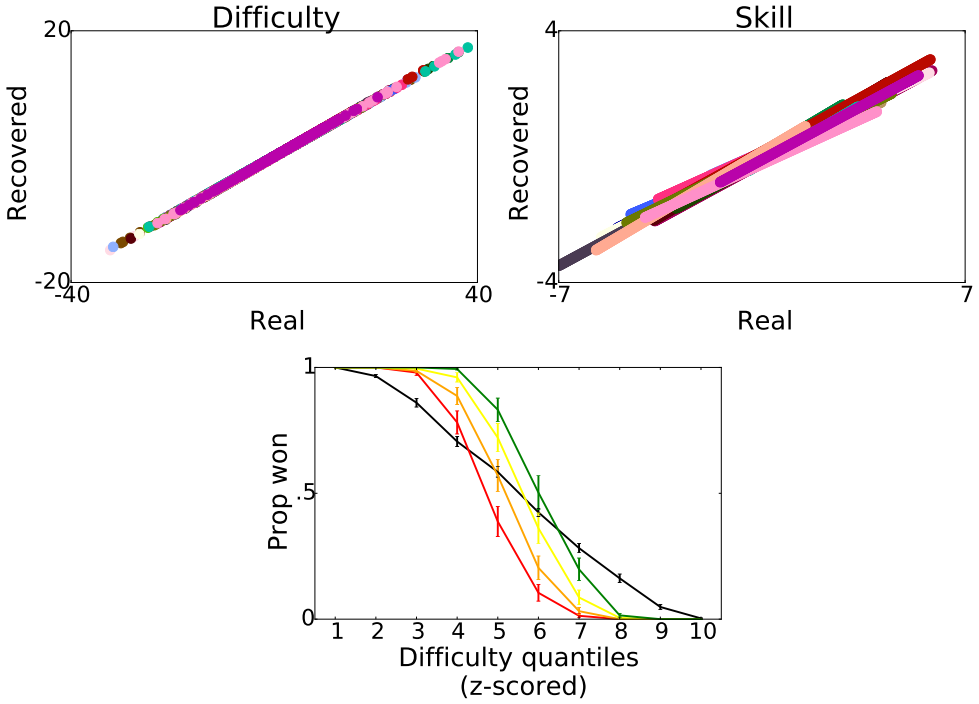


Figure G.4: Difficulty and skill recovery for data in figure G.3. Top difficulty and skill recovery, each color represents one simulated subject. Bottom: relationship between recovered difficulty and the proportion of wins, overall and as a function of recovered skill level; mean $\pm$ s.e.m across subjects; black: overall, red to green: 1^{st} to 4^{th} recovered skill quartiles; recovered difficulty z-scored within subject.

Appendix H

Difficulty measure

As discussed in section 2.3.2, we inferred the weights representing the contributions of different objective task factors to the integrated difficulty measure, $\boldsymbol{w}_d$, separately for each subject, by predicting outcomes for all remaining subjects. We compared two approaches for integrating data from remaining subjects: pooling subjects together and modelling variability between subjects with a hierarchical model. We describe the two approaches below.

Pooling subjects together In this case, in order to find $\boldsymbol{w}_d^{s_0}$ for a particular subject s_0, we considered all trials from all other subjects, $T_{s \neq s_0}$, as if they were generated from the same set of parameters $\boldsymbol{w}_d^{s_0}$, which we inferred as:

$$\boldsymbol{w}_d^{s_0} = \underset{\boldsymbol{w}}{\operatorname{argmin}} \sum_{t \in T_{s \neq s_0}} nll_t(\boldsymbol{w}), \text{ where}$$

$$nll_t(\boldsymbol{w}) = \begin{cases} -\log(\sigma(\boldsymbol{w}^T \boldsymbol{f}_d)) & \text{if } o(t) = 1 \\ -\log(1 - \sigma(\boldsymbol{w}^T \boldsymbol{f}_d)) & \text{if } o(t) = 0 \end{cases}$$

is the negative log likelihood for the outcome of trial t.

Hierarchical model In this case we again used only data from all remaining subjects when inferring $\boldsymbol{w}_d$ for a given subject s_0, but we took into account that parameters for different subjects can be different, and included a model of how they might vary.

Specifically, we used a hierarchical generative model for outcomes, assuming data from each subject $s \neq s_0$ to be generated using a set of parameters

$\boldsymbol{w}_d^s$ and a Gaussian distribution of these parameters across the population:

$$\begin{aligned}
\boldsymbol{w}_d^s &\sim \mathcal{N}(\mu_w, \Sigma_w), \forall \text{ subject } s \neq s_0 \\
o^s(t) &\sim \text{Bernoulli}(\sigma(\boldsymbol{w}_d^{sT} \boldsymbol{f}_d^s(t)), \text{ where} \\
o^s(t) &= \text{outcome of trial t for subject s} \\
\boldsymbol{f}_d^s(t) &= \text{the measured value of } \boldsymbol{f}_d \text{ at trial t for subject s.}
\end{aligned}$$

We used the probabilistic programming language STAN (`https://mc-stan.org/`) to invert this model and obtain samples of the posterior distribution over these population parameters, $p(\mu_w, \Sigma_w | \boldsymbol{f}^s, o^s(t) \forall t, s \neq s_0)$. We then used the expected value of the posterior distribution for μ_w as $\boldsymbol{w}_d^{s_0}$.

The measure thus obtained was highly correlated to the one obtained by pooling data from all subjects: correlation average value 0.87 and s. d. of 0.03.

Appendix I

Performance features: definition, within trial effects, learning effects

We used trial-level measures of performance in our analyses, however there is rich within-trial variability in subjects' behaviour, due to the particular nature of the task: at every moment during a trial, a subject's behaviour is represented by whether they pressed a key or not and, if they did, by their choice of key press; therefore behaviour at the finest-grained level is represented by the key press variable on every frame (every 50 ms).

We labelled the key press variable on every frame as pause (p), correct (c), wrong, but correct for up (wcu), wrong, but correct for previous orientation(wcp) and wrong (w) for all other wrong key presses.

This labelling provides us with a time course of performance within each trial (see figure I.1).

For a qualitative view of the evolution of within-trial performance across trials see figures I.2 and I.3. Subject 02190318's performance improves across trials: in the 2s post rotation internal, they are making fewer pauses and fewer mistakes. The situation is different for subject 02070218, whose number of correct key presses and pattern of mistakes in the 2s interval post rotations doesn't change much in time.

For our analyses we need a coarser, trial-wise summary quantification of performance. We refer to the resulting trial-by-trial measures as performance

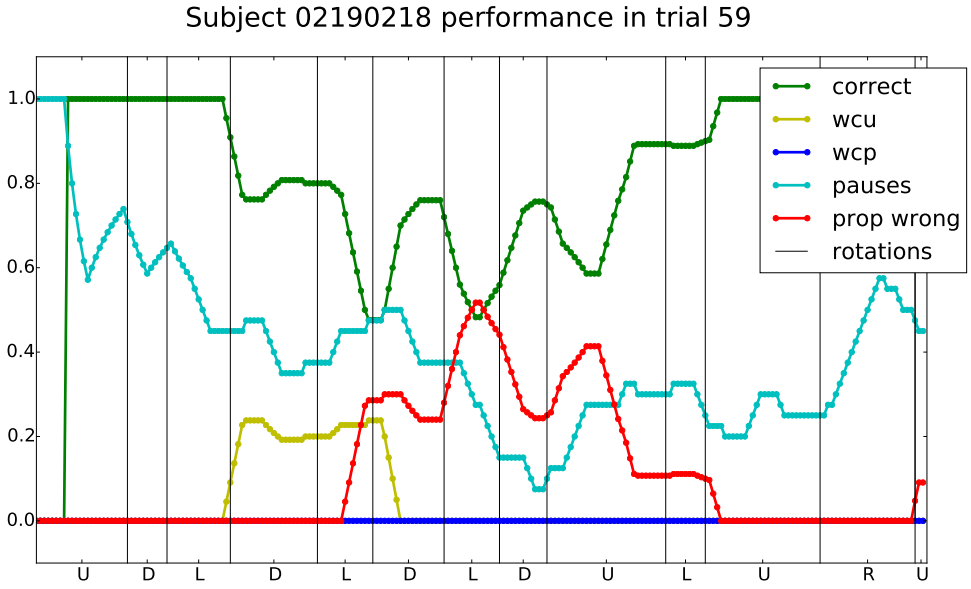


Figure I.1: Example of performance evolution within trial: running averages (2 s smoothing window) of different types of key presses, see legend; vertical black bars represent rotations within the trial, the maze's orientation is labeled as U (up), D (down), L (left), R (right).

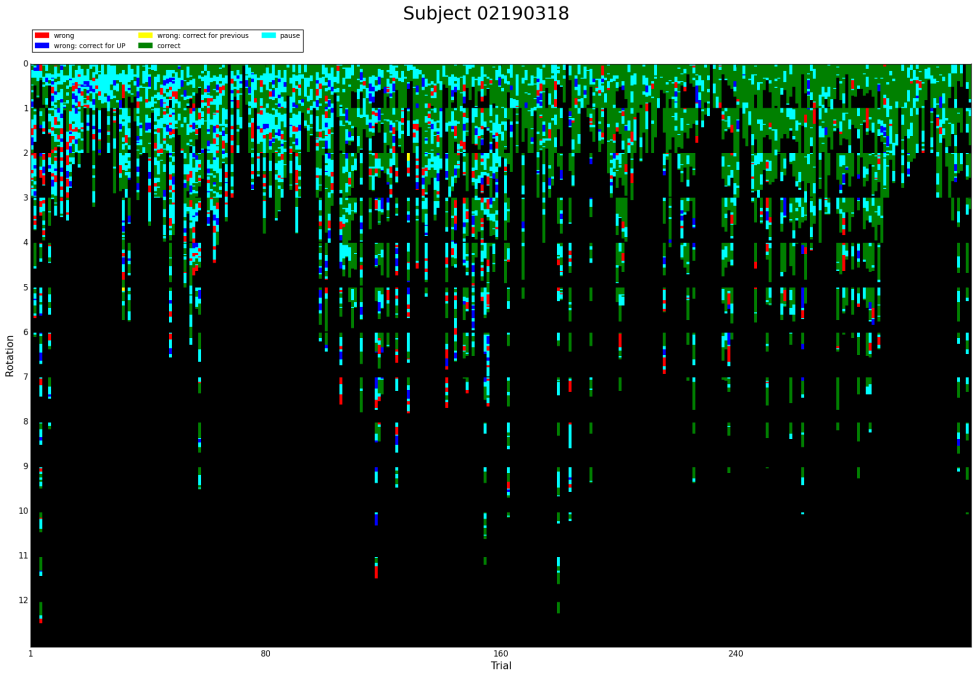


Figure I.2: Post rotation performance evolution, good learner

regressors. We note that despite our intuition about wcp being a relevant label, it turned out that this was an extremely rare type of mistake, and therefore we decided not to include its summary among the performance regressors. For

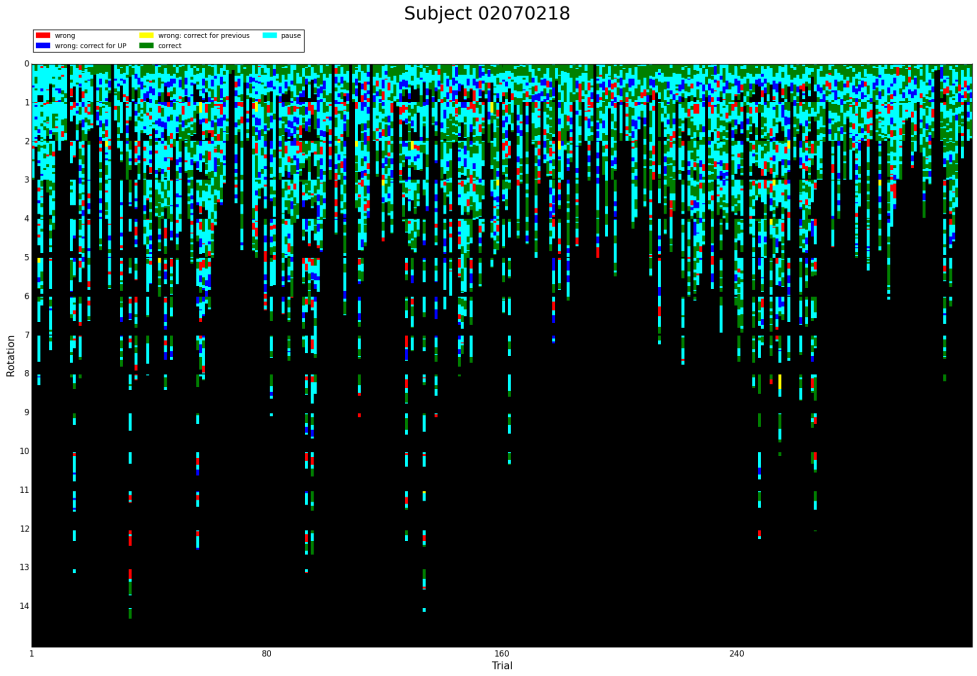


Figure I.3: Post rotation performance evolution subject, bad learner

the same reason, the set of wrong key presses is practically the complement of the set containing correct and wcu key presses, and therefore provides no additional information. We therefore used only the proportion of pauses, the proportion of correct key presses, and the proportion of wrong key presses that would have been correct in the normal UP orientation.

We compared two trial by trial summaries of performance: one in which the proportion of each type of key press was computed across the whole trial, and one in which this proportion was computed considering only the 2s post-rotation interval. In all the analyses described below, there was no difference between using the whole trial and using the post rotation interval versions of the performance regressors. We use the post rotation interval version in all analyses presented.

Appendix J

Model list, skill estimate models

J.1 Purely descriptive models

NOTE: in all the following, t indicates the index of a trial in the whole data set for one subject, while t_s indicates the index of the trial within the session it was recorded in;

$s(t)$ = predicted skill estimate at trial t, and Greek letters represent fitted parameters.

- logarithm shape, ignoring session (model L1):

$$s(t) = \alpha + \beta * \log(t)$$

- logarithm shape + effect of break between sessions, but same parameters for both sessions (model L2):

$$s(t) = \begin{cases} \alpha + \beta * \log(t_s) & \text{if t is in the first session} \\ \alpha + \beta * \log(t_s) + \gamma & \text{if t is in the second session} \end{cases}$$

- logarithm shape, different parameters for the two sessions + effect of break between sessions (model L3):

$$s(t) = \begin{cases} \alpha_1 + \beta_1 * \log(t_s) & \text{if t is in the first session} \\ \alpha_2 + \beta_2 * \log(t_s) + \gamma & \text{if t is in the second session} \end{cases}$$

- sigmoid shape, ignore session (model S1):

$$s(t) = \mu + (\nu - \mu)\sigma(-\beta(x - \alpha))$$

- sigmoid shape + effect of break between sessions, but same parameters for both sessions (model S2):

$$s(t) = \begin{cases} \mu + (\nu - \mu)\sigma(\beta * (t_s - \alpha)) & \text{if t is in the first session} \\ \mu + (\nu - \mu)\sigma(\beta * (t_s - \alpha)) + \gamma & \text{if t is in the second session} \end{cases}$$

- sigmoid shape, different parameters for the two sessions + effect of break between sessions (model S3):

$$s(t) = \begin{cases} \mu_1 + (\nu_1 - \mu_1)\sigma(\beta_1 * (t_s - \alpha_1)) & \text{if t is in the first session} \\ \mu_2 + (\nu_2 - \mu_2)\sigma(\beta 2 * (t_s - \alpha_2)) + \gamma & \text{if t is in the second session} \end{cases}$$

where

$$\sigma(t) = \frac{1}{1 + \exp(-x)}$$

J.2 Rescorla - Wagner models

In all the following model descriptions we use the following notations (greek letters denote additional model parameters):

$s_0 =$ initial value of the internal skill estimate: model parameter

$s_t =$ internal skill estimate after trial t

$r_t =$ skill estimate after trial t

$o_t =$ outcome of trial t $\in \{0, 1\}$

$a_t =$ attribution for trial t $\in \{\text{internal}(1), \text{external}(-1), \text{none}(0)\}$

$t_0^{II} =$ index of first trial of the second session

$\delta_t =$ prediction error at trial t

For most of the models, the naming convention reflects the presence or absence of four orthogonal factors, S (session), A (attribution), O (outcome valence), T (timescale). Other models, such as those including difficulty, follow slightly different naming conventions. All models are fully specified below.

- baseline: this is the basic RW model, which has one learning rate, α, a parameter modelling the effect of the break between sessions, β, and the variance in response noise, σ^2. All other RW models will be variations on this basic one.

$$\delta_t = \begin{cases} o_t - s_{t-1} & \text{if } t \neq t_0^{II} \\ o_t - (s_{t-1} + \beta) & \text{otherwise} \end{cases}$$

$$s_t = \begin{cases} s_{t-1} + \alpha * \delta_t & \text{if } t \neq t_0^{II} \\ s_{t-1} + \beta + \alpha * \delta_t & \text{otherwise} \end{cases}$$

$$r_t = \mathcal{N}(s_t, \sigma^2)$$

- `S___`: baseline + different learning rates for the different sessions:

$$\delta_t = \begin{cases} o_t - s_{t-1} & \text{if } t \neq t_0^{II} \\ o_t - (s_{t-1} + \beta) & \text{otherwise} \end{cases}$$

$$s_t = \begin{cases} s_{t-1} + \alpha(t) * \delta_t & \text{if } t \neq t_0^{II} \\ s_{t-1} + \beta + \alpha(t) * \delta_t & \text{otherwise} \end{cases}, \text{ where } \alpha(t) = \begin{cases} \alpha_1 & \text{if } t < t_0^{II} \\ \alpha_2 & \text{otherwise} \end{cases}$$

$$r_t = \mathcal{N}(s_t, \sigma^2)$$

- `_A__`: baseline + different learning rates for different attributions:

$$\delta_t = \begin{cases} o_t - s_{t-1} & \text{if } t \neq t_0^{II} \\ o_t - (s_{t-1} + \beta) & \text{otherwise} \end{cases}$$

$$s_t = \begin{cases} s_{t-1} + \alpha(t) * \delta_t & \text{if } t \neq t_0^{II} \\ s_{t-1} + \beta + \alpha(t) * \delta_t & \text{otherwise} \end{cases}, \text{ where } \alpha(t) = \begin{cases} \alpha_i & \text{if } a_t = 1 \\ \alpha_e & \text{if } a_t = -1 \\ \alpha_n & \text{otherwise} \end{cases}$$

$$r_t = \mathcal{N}(s_t, \sigma^2)$$

- `__O_`: baseline + different learning rates for wins and losses:

$$\delta_t = \begin{cases} o_t - s_{t-1} & \text{if } t \neq t_0^{II} \\ o_t - (s_{t-1} + \beta) & \text{otherwise} \end{cases}$$

$$s_t = \begin{cases} s_{t-1} + \alpha(t) * \delta_t & \text{if } t \neq t_0^{II} \\ s_{t-1} + \beta + \alpha(t) * \delta_t & \text{otherwise} \end{cases}, \text{ where } \alpha(t) = \begin{cases} \alpha_+ & \text{if } o_t = 1 \\ \alpha_- & \text{if } o_t = 0 \end{cases}$$

$$r_t = \mathcal{N}(s_t, \sigma^2)$$

- `____T`: baseline + outcome impulse effect: local influence of the outcome on the immediately following skill report only, which does not propagate to later trials:

$$\delta_t = \begin{cases} o_t - s_{t-1} & \text{if } t \neq t_0^{II} \\ o_t - (s_{t-1} + \beta) & \text{otherwise} \end{cases}$$

$$s_t = \begin{cases} s_{t-1} + \alpha * \delta_t & \text{if } t \neq t_0^{II} \\ s_{t-1} + \beta + \alpha * \delta_t & \text{otherwise} \end{cases}$$

$$r_t = \mathcal{N}(s_t + \gamma(o_t), \sigma^2), \text{ where } \gamma(o_t) = \begin{cases} \gamma_+ & \text{if } o_t = 1 \\ \gamma_- & \text{otherwise} \end{cases}$$

- `SA__`: baseline + session + attribution:

$$\delta_t = \begin{cases} o_t - s_{t-1} & \text{if } t \neq t_0^{II} \\ o_t - (s_{t-1} + \beta) & \text{otherwise} \end{cases}$$

$$s_t = \begin{cases} s_{t-1} + \alpha(t) * \delta_t & \text{if } t \neq t_0^{II} \\ s_{t-1} + \beta + \alpha(t) * \delta_t & \text{otherwise} \end{cases}, \text{ where } \alpha(t) = \begin{cases} \alpha_i^1 & \text{if } a_t = 1 \text{ and } t < t_0^{II} \\ \alpha_e^1 & \text{if } a_t = -1 \text{ and } t < t_0^{II} \\ \alpha_n^1 & \text{if } a_t = 0 \text{ and } t < t_0^{II} \\ \alpha_i^2 & \text{if } a_t = 1 \text{ and } t \geq t_0^{II} \\ \alpha_e^2 & \text{if } a_t = -1 \text{ and } t \geq t_0^{II} \\ \alpha_n^2 & \text{if } a_t = 0 \text{ and } t \geq t_0^{II} \end{cases}$$

$$r_t = \mathcal{N}(s_t, \sigma^2)$$

- `S_O_`: baseline + session + outcome:

$$\delta_t = \begin{cases} o_t - s_{t-1} & \text{if } t \neq t_0^{II} \\ o_t - (s_{t-1} + \beta) & \text{otherwise} \end{cases}$$

$$s_t = \begin{cases} s_{t-1} + \alpha(t) * \delta_t & \text{if } t \neq t_0^{II} \\ s_{t-1} + \beta + \alpha(t) * \delta_t & \text{otherwise} \end{cases}, \text{ where } \alpha(o_t) = \begin{cases} \alpha_+^1 & \text{if } o_t = 1 \text{ and } t < t_0^{II} \\ \alpha_-^1 & \text{if } o_t = 0 \text{ and } t < t_0^{II} \\ \alpha_+^2 & \text{if } o_t = 1 \text{ and } t \geq t_0^{II} \\ \alpha_-^2 & \text{if } o_t = 0 \text{ and } t \geq t_0^{II} \end{cases}$$

$$r_t = \mathcal{N}(s_t, \sigma^2)$$

- `S__T`: baseline + session + outcome impulse:

$$\delta_t = \begin{cases} o_t - s_{t-1} & \text{if } t \neq t_0^{II} \\ o_t - (s_{t-1} + \beta) & \text{otherwise} \end{cases}$$

$$s_t = \begin{cases} s_{t-1} + \alpha(t) * \delta_t & \text{if } t \neq t_0^{II} \\ s_{t-1} + \beta + \alpha(t) * \delta_t & \text{otherwise} \end{cases}, \text{ where } \alpha(t) = \begin{cases} \alpha_1 & \text{if } t < t_0^{II} \\ \alpha_2 & \text{otherwise} \end{cases}$$

$$r_t = \mathcal{N}(s_t + \gamma(o_t), \sigma^2), \text{ where } \gamma(o_t) = \begin{cases} \gamma_+ & \text{if } o_t = 1 \\ \gamma_- & \text{otherwise} \end{cases}$$

- _AO_: baseline +attribution +outcome:

$$\delta_t = \begin{cases} o_t - s_{t-1} & \text{if } t \neq t_0^{II} \\ o_t - (s_{t-1} + \beta) & \text{otherwise} \end{cases}$$

$$s_t = \begin{cases} s_{t-1} + \alpha(t) * \delta_t & \text{if } t \neq t_0^{II} \\ s_{t-1} + \beta + \alpha(t) * \delta_t & \text{otherwise} \end{cases}, \text{where } \alpha(t) = \begin{cases} \alpha_{i+} & \text{if } o_t = 1 \text{ and } a_t = 1 \\ \alpha_{e+} & \text{if } o_t = 1 \text{ and } a_t = -1 \\ \alpha_{n+} & \text{if } o_t = 1 \text{ and } a_t = 0 \\ \alpha_{i-} & \text{if } o_t = 0 \text{ and } a_t = 1 \\ \alpha_{e-} & \text{if } o_t = 0 \text{ and } a_t = -1 \\ \alpha_{n-} & \text{if } o_t = 0 \text{ and } a_t = 1 = 0 \end{cases}$$

$$r_t = \mathcal{N}(s_t, \sigma^2)$$

- _A_T: baseline + attribution +outcome impulse:

$$\delta_t = \begin{cases} o_t - s_{t-1} & \text{if } t \neq t_0^{II} \\ o_t - (s_{t-1} + \beta) & \text{otherwise} \end{cases}$$

$$s_t = \begin{cases} s_{t-1} + \alpha(t) * \delta_t & \text{if } t \neq t_0^{II} \\ s_{t-1} + \beta + \alpha(t) * \delta_t & \text{otherwise} \end{cases}, \text{ where } \alpha(t) = \begin{cases} \alpha_i & \text{if } a_t = 1 \\ \alpha_e & \text{if } a_t = -1 \\ \alpha_n & \text{otherwise} \end{cases}$$

$$r_t = \mathcal{N}(s_t + \gamma(o_t), \sigma^2), \text{ where } \gamma(o_t) = \begin{cases} \gamma_+ & \text{if } o_t = 1 \\ \gamma_- & \text{otherwise} \end{cases}$$

- __OT: baseline + outcome + outcome impulse:

$$\delta_t = \begin{cases} o_t - s_{t-1} & \text{if } t \neq t_0^{II} \\ o_t - (s_{t-1} + \beta) & \text{otherwise} \end{cases}$$

$$s_t = \begin{cases} s_{t-1} + \alpha(t) * \delta_t & \text{if } t \neq t_0^{II} \\ s_{t-1} + \beta + \alpha(t) * \delta_t & \text{otherwise} \end{cases}, \text{ where } \alpha(o_t) = \begin{cases} \alpha_+ & \text{if } o_t = 1 \\ \alpha_- & \text{if } o_t = 0 \end{cases}$$

$$r_t = \mathcal{N}(s_t + \gamma(o_t), \sigma^2), \text{ where } \gamma(o_t) = \begin{cases} \gamma_+ & \text{if } o_t = 1 \\ \gamma_- & \text{otherwise} \end{cases}$$

- `SAO_`: baseline + session + attribution + outcome.

$$\delta_t = \begin{cases} o_t - s_{t-1} & \text{if } t \neq t_0^{II} \\ o_t - (s_{t-1} + \beta) & \text{otherwise} \end{cases}$$

$$s_t = \begin{cases} s_{t-1} + \alpha(t) * \delta_t & \text{if } t \neq t_0^{II} \\ s_{t-1} + \beta + \alpha(t) * \delta_t & \text{otherwise} \end{cases}, \text{ where } \alpha(t) = \begin{cases} \alpha_{i+}^1 & \text{if } o_t = 1, a_t = 1 \text{ and } t < t_0^{II} \\ \alpha_{e+}^1 & \text{if } o_t = 1, a_t = -1 \text{ and } t < t_0^{II} \\ \alpha_{n+}^1 & \text{if } o_t = 1, a_t = 0 \text{ and } t < t_0^{II} \\ \alpha_{i-}^1 & \text{if } o_t = 0, a_t = 1 \text{ and } t < t_0^{II} \\ \alpha_{e-}^1 & \text{if } o_t = 0, a_t = -1 \text{ and } t < t_0^{II} \\ \alpha_{n-}^1 & \text{if } o_t = 0, a_t = 0 \text{ and } t < t_0^{II} \\ \alpha_{i+}^2 & \text{if } o_t = 1, a_t = 1 \text{ and } t \geq t_0^{II} \\ \alpha_{e+}^2 & \text{if } o_t = 1, a_t = -1 \text{ and } t \geq t_0^{II} \\ \alpha_{n+}^2 & \text{if } o_t = 1, a_t = 0 \text{ and } t \geq t_0^{II} \\ \alpha_{i-}^2 & \text{if } o_t = 0, a_t = 1 \text{ and } t \geq t_0^{II} \\ \alpha_{e-}^2 & \text{if } o_t = 0, a_t = -1 \text{ and } t \geq t_0^{II} \\ \alpha_{n-}^2 & \text{if } o_t = 0, a_t = 0 \text{ and } t \geq t_0^{II} \end{cases}$$

$$r_t = \mathcal{N}(s_t, \sigma^2)$$

- `SA_T`: baseline + session + attribution + outcome impulse:

$$\delta_t = \begin{cases} o_t - s_{t-1} & \text{if } t \neq t_0^{II} \\ o_t - (s_{t-1} + \beta) & \text{otherwise} \end{cases}$$

$$s_t = \begin{cases} s_{t-1} + \alpha(t) * \delta_t & \text{if } t \neq t_0^{II} \\ s_{t-1} + \beta + \alpha(t) * \delta_t & \text{otherwise} \end{cases}, \text{where } \alpha(t) = \begin{cases} \alpha_i^1 & \text{if } a_t = 1 \text{ and } t < t_0^{II} \\ \alpha_e^1 & \text{if } a_t = -1 \text{ and } t < t_0^{II} \\ \alpha_n^1 & \text{if } a_t = 0 \text{ and } t < t_0^{II} \\ \alpha_i^2 & \text{if } a_t = 1 \text{ and } t \geq t_0^{II} \\ \alpha_e^2 & \text{if } a_t = -1 \text{ and } t \geq t_0^{II} \\ \alpha_n^2 & \text{if } a_t = 0 \text{ and } t \geq t_0^{II} \end{cases}$$

$$r_t = \mathcal{N}(s_t + \gamma(o_t), \sigma^2), \text{ where } \gamma(o_t) = \begin{cases} \gamma_+ & \text{if } o_t = 1 \\ \gamma_- & \text{otherwise} \end{cases}$$

- `S_OT`: baseline + session + outcome + outcome impulse:

$$\delta_t = \begin{cases} o_t - s_{t-1} & \text{if } t \neq t_0^{II} \\ o_t - (s_{t-1} + \beta) & \text{otherwise} \end{cases}$$

$$s_t = \begin{cases} s_{t-1} + \alpha(t) * \delta_t & \text{if } t \neq t_0^{II} \\ s_{t-1} + \beta + \alpha(t) * \delta_t & \text{otherwise} \end{cases}, \text{ where } \alpha(t) = \begin{cases} \alpha_+^1 & \text{if } o_t = 1 \text{ and } t < t_0^{II} \\ \alpha_-^1 & \text{if } o_t = 0 \text{ and } t < t_0^{II} \\ \alpha_+^2 & \text{if } o_t = 1 \text{ and } t \geq t_0^{II} \\ \alpha_-^2 & \text{if } o_t = 0 \text{ and } t \geq t_0^{II} \end{cases}$$

$$r_t = \mathcal{N}(s_t + \gamma(o_t), \sigma^2), \text{ where } \gamma(o_t) = \begin{cases} \gamma_+ & \text{if } o_t = 1 \\ \gamma_- & \text{otherwise} \end{cases}$$

- `_AOT`: baseline + attribution + outcome + outcome impulse:

$$\delta_t = \begin{cases} o_t - s_{t-1} & \text{if } t \neq t_0^{II} \\ o_t - (s_{t-1} + \beta) & \text{otherwise} \end{cases}$$

$$s_t = \begin{cases} s_{t-1} + \alpha(t) * \delta_t & \text{if } t \neq t_0^{II} \\ s_{t-1} + \beta + \alpha(t) * \delta_t & \text{otherwise} \end{cases}, \text{ where } \alpha(t) = \begin{cases} \alpha_{i+} & \text{if } o_t = 1, a_t = 1 \\ \alpha_{e+} & \text{if } o_t = 1, a_t = -1 \\ \alpha_{n+} & \text{if } o_t = 1, a_t = 0 \\ \alpha_{i-} & \text{if } o_t = 0, a_t = 1 \\ \alpha_{e-} & \text{if } o_t = 0, a_t = -1 \\ \alpha_{n-} & \text{if } o_t = 0, a_t = 0 \end{cases}$$

$$r_t = \mathcal{N}(s_t + \gamma(o_t), \sigma^2), \text{ where } \gamma(o_t) = \begin{cases} \gamma_+ & \text{if } o_t = 1 \\ \gamma_- & \text{otherwise} \end{cases}$$

- `SAOT`: baseline + session + attribution + outcome + outcome impulse:

$$\delta_t = \begin{cases} o_t - s_{t-1} & \text{if } t \neq t_0^{II} \\ o_t - (s_{t-1} + \beta) & \text{otherwise} \end{cases}, s_t = \begin{cases} s_{t-1} + \alpha(t) * \delta_t & \text{if } t \neq t_0^{II} \\ s_{t-1} + \beta + \alpha(t) * \delta_t & \text{otherwise} \end{cases},$$

$$\text{where } \alpha(t) = \begin{cases} \alpha_{i+}^1 & \text{if } o_t = 1, a_t = 1 \text{ and } t < t_0^{II} \\ \alpha_{e+}^1 & \text{if } o_t = 1, a_t = -1 \text{ and } t < t_0^{II} \\ \alpha_{n+}^1 & \text{if } o_t = 1, a_t = 0 \text{ and } t < t_0^{II} \\ \alpha_{i-}^1 & \text{if } o_t = 0, a_t = 1 \text{ and } t < t_0^{II} \\ \alpha_{e-}^1 & \text{if } o_t = 0, a_t = -1 \text{ and } t < t_0^{II} \\ \alpha_{n-}^1 & \text{if } o_t = 0, a_t = 0 \text{ and } t < t_0^{II} \\ \alpha_{i+}^2 & \text{if } o_t = 1, a_t = 1 \text{ and } t \geq t_0^{II} \\ \alpha_{e+}^2 & \text{if } o_t = 1, a_t = -1 \text{ and } t \geq t_0^{II} \\ \alpha_{n+}^2 & \text{if } o_t = 1, a_t = 0 \text{ and } t \geq t_0^{II} \\ \alpha_{i-}^2 & \text{if } o_t = 0, a_t = 1 \text{ and } t \geq t_0^{II} \\ \alpha_{e-}^2 & \text{if } o_t = 0, a_t = -1 \text{ and } t \geq t_0^{II} \\ \alpha_{n-}^2 & \text{if } o_t = 0, a_t = 0 \text{ and } t \geq t_0^{II} \end{cases}$$

$$r_t = \mathcal{N}(s_t + \gamma(o_t), \sigma^2), \text{ where } \gamma(o_t) = \begin{cases} \gamma_+ & \text{if } o_t = 1 \\ \gamma_- & \text{otherwise} \end{cases}$$

- RWD1: in this model the difficulty of the current trial contributes to the outcome prediction in δ:

$$\delta_t = \begin{cases} o_t - (0.5 + 0.5 * (s_{t-1} - d_t)) & \text{if } t \neq t_0^{II} \\ o_t - (0.5 + 0.5 * (s_{t-1} + \beta - d_t)) & \text{otherwise} \end{cases}$$

$$s_t = \begin{cases} s_{t-1} + \alpha * \delta_t & \text{if } t \neq t_0^{II} \\ s_{t-1} + \beta + \alpha * \delta_t & \text{otherwise} \end{cases}$$

$$r_t = \mathcal{N}(s_t, \sigma^2)$$

- RWD2: this a model in which difficulty $d(t)$ modulates the prediction error effect on the internal skill update, as follows:

$$\delta_t = \begin{cases} o_t - s_{t-1} & \text{if } t \neq t_0^{II} \\ o_t - (s_{t-1} + \beta) & \text{otherwise} \end{cases}$$

$$s_t = \begin{cases} s_{t-1} + \alpha * \delta_t * d(t) & \text{if } t \neq t_0^{II} \\ s_{t-1} + \beta + \alpha * \delta_t * d(t) & \text{otherwise} \end{cases}$$

$$r_t = \mathcal{N}(s_t, \sigma^2)$$

- RWD3: this is a model in which the difficulty $d(t)$ and outcome of the current trial modulate the prediction error effect on the internal skill update, as follows:

$$\delta_t = \begin{cases} o_t - s_{t-1} & \text{if } t \neq t_0^{II} \\ o_t - (s_{t-1} + \beta) & \text{otherwise} \end{cases}$$

$$s_t = \begin{cases} s_{t-1} + \alpha * \delta_t * df(t) & \text{if } t \neq t_0^{II} \\ s_{t-1} + \beta + \alpha * \delta_t * df(t) & \text{otherwise} \end{cases}, \text{ where } df(t) = \begin{cases} d(t) & \text{if } o_t = 1 \\ 1 - d(t) & \text{if } o_t = 0 \end{cases},$$

$$r_t = \mathcal{N}(s_t, \sigma^2)$$

- __○_+diff: this is a different model of the interaction of difficulty and outcome in modulating the internal skill update; difficulty $d(t)$ mod-

ulates the prediction error effect on the internal skill update equally regardless of outcome, but learning rate can be different for wins and losses:

$$\delta_t = \begin{cases} o_t - s_{t-1} & \text{if } t \neq t_0^{II} \\ o_t - (s_{t-1} + \beta) & \text{otherwise} \end{cases}$$

$$s_t = \begin{cases} s_{t-1} + \alpha(t) * \delta_t * d(t) & \text{if } t \neq t_0^{II} \\ s_{t-1} + \beta + \alpha(t) * \delta_t * d(t) & \text{otherwise} \end{cases}, \text{ where } \alpha(t) = \begin{cases} \alpha_+ & \text{if } o_t = 1 \\ \alpha_- & \text{if } o_t = 0 \end{cases}$$

$$r_t = \mathcal{N}(s_t, \sigma^2)$$

- _A__+diff: this is similar to __O_+diff, but difficulty $d(t)$ and attribution, rather than outcome, modulate the internal skill update

$$\delta_t = \begin{cases} o_t - s_{t-1} & \text{if } t \neq t_0^{II} \\ o_t - (s_{t-1} + \beta) & \text{otherwise} \end{cases}$$

$$s_t = \begin{cases} s_{t-1} + \alpha(t) * \delta_t * d(t) & \text{if } t \neq t_0^{II} \\ s_{t-1} + \beta + \alpha(t) * \delta_t * d(t) & \text{otherwise} \end{cases}, \text{ where } \alpha(t) = \begin{cases} \alpha_i & \text{if } a_t = 1 \\ \alpha_e & \text{if } a_t = -1 \\ \alpha_n & \text{otherwise} \end{cases}$$

$$r_t = \mathcal{N}(s_t, \sigma^2)$$

- RWD4: this is a model in which the learning rate is a linear function of the difficulty $d(t)$:

$$\delta_t = \begin{cases} o_t - s_{t-1} & \text{if } t \neq t_0^{II} \\ o_t - (s_{t-1} + \beta) & \text{otherwise} \end{cases}$$

$$s_t = \begin{cases} s_{t-1} + (\varepsilon + \phi * d(t)) * \delta_t & \text{if } t \neq t_0^{II} \\ s_{t-1} + \beta + (\varepsilon + \phi * d(t)) * \delta_t & \text{otherwise} \end{cases}$$

$$r_t = \mathcal{N}(s_t, \sigma^2)$$

- RWD4+O: this is similar to RWD4, but the linear relationship between

difficulty and learning rate is allowed to be different for wins and losses:

$$\delta_t = \begin{cases} o_t - s_{t-1} & \text{if } t \neq t_0^{II} \\ o_t - (s_{t-1} + \beta) & \text{otherwise} \end{cases}$$

$$s_t = \begin{cases} s_{t-1} + \alpha(t, d(t)) * \delta_t & \text{if } t \neq t_0^{II} \\ s_{t-1} + \beta + \alpha(t, d(t)) * \delta_t & \text{otherwise} \end{cases},$$

$$\text{where } \alpha(t, d(t)) = \begin{cases} \varepsilon_+ + \phi_+ * d(t) & \text{if } o_t = 1 \\ \varepsilon_- + \phi_- * d(t) & \text{otherwise} \end{cases}$$

$$r_t = \mathcal{N}(s_t, \sigma^2)$$

- RWD5: this is a model in which the relationship between difficulty and learning rate is quadratic:

$$\delta_t = \begin{cases} o_t - s_{t-1} & \text{if } t \neq t_0^{II} \\ o_t - (s_{t-1} + \beta) & \text{otherwise} \end{cases}$$

$$s_t = \begin{cases} s_{t-1} + (\varepsilon + \phi * d(t) + \theta * d(t)^2) * \delta_t & \text{if } t \neq t_0^{II} \\ s_{t-1} + \beta + (\varepsilon + \phi * d(t) + \theta * d(t)^2) * \delta_t & \text{otherwise} \end{cases}$$

$$r_t = \mathcal{N}(s_t, \sigma^2)$$

- RWD5+O: this is an augmentation of RW16, in which the coefficients of the quadratic relationship between difficulty and learning rate are allowed to be different for different outcomes:

$$\delta_t = \begin{cases} o_t - s_{t-1} & \text{if } t \neq t_0^{II} \\ o_t - (s_{t-1} + \beta) & \text{otherwise} \end{cases}$$

$$s_t = \begin{cases} s_{t-1} + \alpha(t) * \delta_t & \text{if } t \neq t_0^{II} \\ s_{t-1} + \beta + \alpha(t) * \delta_t & \text{otherwise} \end{cases},$$

$$\text{where } \alpha(t) = \begin{cases} \varepsilon_+ + \phi_+ * d(t) + \theta_+ * d(t)^2 & \text{if } o_t = 1 \\ \varepsilon_- + \phi_- * d(t) + \theta_- * d(t)^2 & \text{otherwise} \end{cases}$$

$$r_t = \mathcal{N}(s_t, \sigma^2)$$

- RWD6: in this model the prediction part of the prediction error is a combination of the current value of the internal skill estimate and a filtered history of encountered difficulties, $fd(t)$:

$$\begin{aligned} fd(0) &= d(0) \\ fd(t) &= \varepsilon * fd(t-1) + (1-\varepsilon) * d(t) \\ \delta_t &= o_t - (0.5 * (s_{t-1} - fd(t)) + 0.5) \\ s_t &= \begin{cases} s_{t-1} + \alpha * \delta_t & \text{if } t \neq t_0^{II} \\ s_{t-1} + \beta + \alpha * \delta_t & \text{otherwise} \end{cases} \\ r_t &= \mathcal{N}(s_t, \sigma^2) \end{aligned}$$

- RWD6+T: this is model RWD6 with the addition of a local effect of outcome on the skill estimate only:

$$\begin{aligned} fd(0) &= d(0) \\ fd(t) &= \varepsilon * fd(t-1) + (1-\varepsilon) * d(t) \\ \delta_t &= o_t - (0.5 * (s_{t-1} - fd(t)) + 0.5) \\ s_t &= \begin{cases} s_{t-1} + \alpha * \delta_t & \text{if } t \neq t_0^{II} \\ s_{t-1} + \beta + \alpha * \delta_t & \text{otherwise} \end{cases} \\ r_t &= \mathcal{N}(s_t + \gamma(o_t), \sigma^2), \text{ where } \gamma(o_t) = \begin{cases} \gamma_+ & \text{if } o_t = 1 \\ \gamma_- & \text{otherwise} \end{cases} \end{aligned}$$

- RWD6+S: this is model RWD6 augmented with different learning rates for the two sessions:

$$
\begin{aligned}
fd(0) &= d(0) \\
fd(t) &= \varepsilon * fd(t-1) + (1-\varepsilon) * d(t) \\
\delta_t &= o_t - (0.5 * (s_{t-1} - fd(t)) + 0.5) \\
s_t &= \begin{cases} s_{t-1} + \alpha_1 * \delta_t & \text{if } t < t_0^2 \\ s_{t-1} + \beta + \alpha_2 * \delta_t & \text{if } t = t_0^{II} \\ s_{t-1} + \alpha_2 * \delta_t & \text{if } t > t_0^{II} \end{cases} \\
r_t &= \mathcal{N}(s_t, \sigma^2)
\end{aligned}
$$

- RWD6+ST: this is model RWD6 with different parameters for the two sessions and augmented with the addition of a local effect of outcome on the skill estimate only:

$$
r_t = \mathcal{N}(s_t + \gamma(o_t), \sigma^2), \text{ where } \gamma(o_t) = \begin{cases} \gamma_+ & \text{if } o_t = 1 \\ \gamma_- & \text{otherwise} \end{cases}
$$

J.3 BIC score computation for curve fitting models

The BIC score is defined as

$$
k \log(N) - 2 \log(\hat{L}),
$$

where k is the number of parameters, N the number of trials that they contribute to explaining and $\hat{L}$ the maximum likelihood of the data under the model.

Curve fitting models assume responses are independent conditioned on parameters, and distributed normally around predictions, with a fixed and un-

known standard deviation σ. The BIC therefore becomes

$$\begin{aligned} BIC &= k\log(N) - 2\Sigma_{i=1}^{N}\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y_i-\hat{y}_i)^2}{2\sigma^2}}\right) \\ &= k\log(N) - 2\Sigma_{i=1}^{N}\left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_i-\hat{y}_i)^2}{2\sigma^2}\right) \\ &= k\log(N) + N\log(2\pi\sigma^2) + \frac{1}{\sigma^2}\Sigma_{i=1}^{N}(y_i-\hat{y}_i)^2 \end{aligned}$$

Keeping only the parts that differ between models we are left with

$$k\log(N) + N\log(\sigma^2) + \frac{1}{\sigma^2}\Sigma_{i=1}^{N}(y_i-\hat{y}_i)^2.$$

If we approximate σ^2 with $\frac{\Sigma_{i=1}^{N}(y_i-\hat{y}_i)^2}{N}$, then the above becomes:

$$k\log(N) + N\log\left(\frac{\Sigma_{i=1}^{N}(y_i-\hat{y}_i)^2}{N}\right) + \frac{N}{\Sigma_{i=1}^{N}(y_i-\hat{y}_i)^2}\Sigma_{i=1}^{N}(y_i-\hat{y}_i)^2$$

$$k\log(N) - N\log(N) + N\log\left(\Sigma_{i=1}^{N}(y_i-\hat{y}_i)^2\right) + N,$$

which, if we again keep only the terms that are different between models, becomes

$$k\log(N) + N\log\left(\Sigma_{i=1}^{N}(y_i-\hat{y}_i)^2\right).$$

The r^2 score obtained from curve fitting models is

$$r^2 = 1 - \frac{\Sigma_{i=1}^{N}(y_i-\hat{y}_i)^2}{\Sigma_{i=1}^{N}(y_i-\bar{y})^2},$$

therefore

$$\Sigma_{i=1}^{N}(y_i-\hat{y}_i)^2 = (1-r^2)\Sigma_{i=1}^{N}(y_i-\bar{y})^2 = (1-r^2)SS,$$

where SS does not depend on the model.

Plugging the expression for $\Sigma_{i=1}^{N}(y_i-\hat{y}_i)^2$ in the reduced formula for BIC

obtained above we obtain:

$$\begin{aligned} & k\log(N) + N\log\left(\Sigma_{i=1}^{N}(y_i - \hat{y}_i)^2\right) = \\ & k\log(N) + N\log\left((1-r^2)SS\right) = \\ & k\log(N) + N\log\left(1-r^2\right) + N\log\left(SS\right). \end{aligned} \tag{J.1}$$

As SS does not depend on the model, the last term can also be dropped for the purpose of model comparison. Therefore the reduced approximation to the BIC score becomes

$$k\log(N) + N\log\left(1-r^2\right),$$

which can be computed directly from r^2 scores for curve fitting models.

Appendix K

Observing the observer models. Posterior approximations and updates

For all the models, the propagation of subjects' belief about the underlying skill is a filtering process, which involves computing the update after every trial, therefore producing the posterior belief conditioned on the observations of that trial, and propagating this belief through the skill evolution function, therefore obtaining the prior belief before observing the following trial. Due to the simple nature of the skill evolution functions that we used, and to the fact that the belief distributions to be propagated are Gaussian approximations(see below), propagating beliefs through the skill evolution functions is straightforward. However computing the posterior is challenging: in all but very few cases - such as when both the prior and the likelihood are Gaussians- computing the posterior distribution involves intractable integrals. The same is valid in our case, since the observation likelihood is Bernoulli, not Gaussian; we therefore approximated the true posterior with Gaussian distributions, as detailed below.

Written in the most general case, where θ denotes the parameter and o

denotes the observation, the Bayesian posterior computation is

$$p(\theta|o) = \frac{p(o,\theta)}{p(o)} = \frac{p(\theta) * p(o|\theta)}{p(o)} = \frac{p(\theta) * p(o|\theta)}{\int d\theta p(\theta) * p(o|\theta)}$$

One simple approach is to use a Laplace approximation: the issue appears because of the need to compute the normalising constant of the numerator in the above fraction, seen as a function of θ; the solution consists in approximating this function with a simple one - a Gaussian bump situated at a maximum of the original function; the Gaussian bump is obtained by making a 2nd order approximation of the log of the numerator around a maximum point. The resulting approximation is therefore:

$$\begin{aligned} &\mathcal{N}(\theta|\mu_{posterior}, \Sigma_{posterior}), \text{ where} \\ \mu_{posterior} &= \arg\min(-\log(p(\theta) * p(o|\theta))) \\ \Sigma_{posterior} &= H^{-1}, \text{ where} \\ H &= \text{Hessian of the objective function } p(\theta) * p(o|\theta) \\ &\quad \text{at the minimum point } \mu_{posterior}. \end{aligned}$$

While this approach is simple theoretically and easy to implement, it has a number of drawbacks Bishop (2006) the most relevant to us here being its high computational cost: the approximation relies on optimisation, and in our case, where the approximation is needed for belief updating after every trial, this is particularly costly.

Based on the approach of Mathys at al Mathys et al. (2011) , we decided to use a different Gaussian approximation: instead of performing an optimisation to find a maximum of the unnormalised posterior, we used the current estimate of the parameter as the point around which the quadratic approximation is made.

What we need is an approximation of $p(s_t|o_1,o_2,...o_t)$.

$$\begin{aligned} p(s_t|o_1,o_2,...o_t) &= \frac{p(s_t,o_t|o_1,..,o_{t-1})}{p(o_t|o_1,...o_{t-1})} \\ &= \frac{p(s_t|o_1,..,o_{t-1})p(o_t|s_t,o_1,...o_{t-1})}{p(o_t|o_1,...o_{t-1})} \\ &= \frac{p(s_t|o_1,..,o_{t-1})p(o_t|s_t)}{p(o_t|o_1,...o_{t-1})}, \end{aligned}$$

because o_t is independent of previous outcomes given s_t.

We want to approximate $p(s_t|o_1,o_2,...o_t)$ with a Gaussian, which is equivalent to approximating its logarithm with a quadratic function:

$$\begin{aligned} \log\frac{p(s_t|o_1,..,o_{t-1})p(o_t|s_t)}{p(o_t|o_1,...o_{t-1})} &\approx -\frac{1}{2}\log(2\pi\sigma_{new}^2) - \frac{(s_t-\mu_{new})^2}{2\sigma_{new}^2} \\ \log(p(s_t|o_1,..,o_{t-1}))+ & \\ \log(p(o_t|s_t))- & \\ \log(p(o_t|o_1,...o_{t-1})) &\approx -\frac{1}{2}\log(2\pi\sigma_{new}^2) - \frac{(s_t-\mu_{new})^2}{2\sigma_{new}^2} \end{aligned}$$

The third term on the left hand side is a constant w.r.t s_t. The first term is already quadratic, since it is the log of the propagated Gaussian density, $p(s_t|o_1,...o_{t-1}) = \mathcal{N}(s_t|\mu_{old},\sigma_{old}^2)$. We approximate the middle term with a quadratic expression in s_t by keeping only the terms up to second order in the Taylor expansion of $\log(p(o_t|s_t))$ around μ_{old}.

If observed outcome is win

If $o_t = 1$, then

$$p(o_t|s_t) = \sigma(\beta(d_t - s_t))$$

and the 2nd order approximation around μ_{old} is

$$\begin{aligned} \log(p(1|s_t)) \approx &\log(p(1|\mu_{old})) \\ &-\beta(1-\sigma(\beta(d_t-\mu_{old}))) \\ &-\frac{1}{2}\beta^2(1-\sigma(\beta(d_t-\mu_{old})))\sigma(\beta(d_t-\mu_{old}))(s_t-\mu_{old})^2 \end{aligned}$$

The approximation above then becomes

$$
\begin{aligned}
&-\frac{1}{2}\log(2\pi\sigma_{old}^2) - \frac{(s_t-\mu_{old})^2}{2\sigma_{old}^2} + \log(p(1|\mu_{old})) - \beta(1-\sigma(\beta(d_t-\mu_{old})))\\
&-\frac{1}{2}\beta^2(1-\sigma(\beta(d_t-\mu_{old})))\sigma(\beta(d_t-\mu_{old}))(s_t-\mu_{old})^2 + \log p(o_t|o_1,...o_{t-1})\\
&= -\frac{1}{2}\log(2\pi\sigma_{new}^2) - \frac{(s_t-\mu_{new})^2}{2\sigma_{new}^2}
\end{aligned}
$$

and what is needed is to identify μ_{new} and σ_{new}^2.

For this we use Mathys et al's Mathys et al. (2011) observations about situations where, for a quadratic function, f,

$f(x) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(s_t-\mu)^2}{2\sigma^2}$.

In such a situation, the following hold:

$$
\begin{aligned}
\sigma^2 &= -\frac{1}{\frac{d^2f}{dx}} \text{ and}\\
\mu &= x_0 - \frac{\frac{df}{dx}(x_0)}{\frac{d^2f}{dx}(x_0)} \text{ for any } x_0.
\end{aligned}
$$

The first is obtained by deriving both sides of the equations twice. The second is a general property of a second order function, which allows to find its argmax in one step from any starting point. In the present case, applying these two equations by taking $x_0 = \mu_{old}$ leads to the updates:

$$
\begin{aligned}
\sigma_{new}^2 &= \frac{1}{\frac{1}{\sigma_{old}^2} + \beta^2(1-\sigma(\beta(d_t-\mu_{old})))\sigma(\beta(d_t-\mu_{old}))}\\
\mu_{new} &= \mu_{old} - \frac{\beta(1-\sigma(\beta(d_t-\mu_{old})))}{\frac{1}{\sigma_{old}^2} + \beta^2(1-\sigma(\beta(d_t-\mu_{old})))\sigma(\beta(d_t-\mu_{old}))}
\end{aligned}
$$

If observed outcome is loss

If $o_t = 0$, then

$$p(o_t|s_t) = 1 - \sigma(\beta(d_t - s_t))$$

and the 2nd order approximation around μ_{old} is

$$\begin{aligned}\log(p(1|s_t)) \approx \log(p(1|\mu_{old})) \\ + \beta\sigma(\beta(d_t - \mu_{old})) \\ - \frac{1}{2}\beta^2(1 - \sigma(\beta(d_t - \mu_{old})))\sigma(\beta(d_t - \mu_{old}))(s_t - \mu_{old})^2\end{aligned}$$

The approximation above then becomes

$$\begin{aligned}&-\frac{1}{2}\log(2\pi\sigma_{old}^2) - \frac{(s_t - \mu_{old})^2}{2\sigma_{old}^2} + \log(p(1|\mu_{old})) + \beta\sigma(\beta(d_t - \mu_{old})) \\ &-\frac{1}{2}\beta^2(1 - \sigma(\beta(d_t - \mu_{old})))\sigma(\beta(d_t - \mu_{old}))(s_t - \mu_{old})^2 + \log p(o_t|o_1, ...o_{t-1}) \\ &= -\frac{1}{2}\log(2\pi\sigma_{new}^2) - \frac{(s_t - \mu_{new})^2}{2\sigma_{new}^2}.\end{aligned}$$

Using again the above equations to identify identify μ_{new} and σ_{new} leads to the updates:

$$\begin{aligned}\sigma_{new}^2 &= \frac{1}{\frac{1}{\sigma_{old}^2} + \beta^2(1 - \sigma(\beta(d_t - \mu_{old})))\sigma(\beta(d_t - \mu_{old}))} \\ \mu_{new} &= \mu_{old} + \frac{\beta\sigma(\beta(d_t - \mu_{old}))}{\frac{1}{\sigma_{old}^2} + \beta^2(1 - \sigma(\beta(d_t - \mu_{old})))\sigma(\beta(d_t - \mu_{old}))}\end{aligned}$$

To sum up, updates can be approximated as follows:

$$\begin{aligned}\sigma_{new}^2 &= \frac{1}{\frac{1}{\sigma_{old}^2} + \beta^2(1 - \sigma(\beta(d_t - \mu_{old})))\sigma(\beta(d_t - \mu_{old}))} \\ \mu_{new} &= \begin{cases} \mu_{old} - \frac{\beta(1 - \sigma(\beta(d_t - \mu_{old})))}{\frac{1}{\sigma_{old}^2} + \beta^2(1 - \sigma(\beta(d_t - \mu_{old})))\sigma(\beta(d_t - \mu_{old}))} & \text{if } o_t = 1 \\ \mu_{old} + \frac{\beta\sigma(\beta(d_t - \mu_{old}))}{\frac{1}{\sigma_{old}^2} + \beta^2(1 - \sigma(\beta(d_t - \mu_{old})))\sigma(\beta(d_t - \mu_{old}))} & \text{otherwise,} \end{cases}\end{aligned} \tag{K.1}$$

where

$$o_t \sim Bernoulli(\sigma(d_t - s_t))$$

$$\mathcal{N}(\mu_{new}, \sigma^2_{new}) \approx p(s_t|o_1, o_2, ...o_t)$$

$$\mathcal{N}(\mu_{old}, \sigma^2_{old}) \approx p(s_t|o_1, o_2, ...o_{t-1})$$

$$\sigma(x) = \frac{1}{1+\exp(-x)} \text{ and}$$

$$\beta < 0 = \text{ fixed slope of the generative model for the outcome.}$$

Adding a modulation of the update according to attribution or outcome leads to the following updates:

$$\sigma^2_{new} = \frac{1}{\frac{1}{\sigma^2_{old}} + \alpha_a * \beta^2(1-\sigma(\beta(d_t-\mu_{old})))\sigma(\beta(d_t-\mu_{old}))}$$

$$\mu_{new} = \begin{cases} \mu_{old} - \alpha * \frac{\beta(1-\sigma(\beta(d_t-\mu_{old})))}{\frac{1}{\sigma^2_{old}}+\beta^2(1-\sigma(\beta(d_t-\mu_{old})))\sigma(\beta(d_t-\mu_{old}))} & \text{if } o_t = 1 \\ \mu_{old} + \alpha * \frac{\beta\sigma(\beta(d_t-\mu_{old}))}{\frac{1}{\sigma^2_{old}}+\beta^2(1-\sigma(\beta(d_t-\mu_{old})))\sigma(\beta(d_t-\mu_{old}))} & \text{otherwise,} \end{cases} \quad \text{(K.2)}$$

where α has different values for the different attributions

$$\alpha = \begin{cases} 1 & \text{if no attribution exists for trial} \\ \alpha_i & \text{if trial attributed internally} \\ \alpha_e & \text{if trial attributed externally} \end{cases}$$

or different values for the different outcomes,

$$\alpha = \begin{cases} \alpha_+ & \text{if trial won} \\ \alpha_- & \text{if trial lost} \end{cases}$$

according to the model.

Appendix L

Correlations between features of interest

Figure L.1 shows correlations between the features of interest.

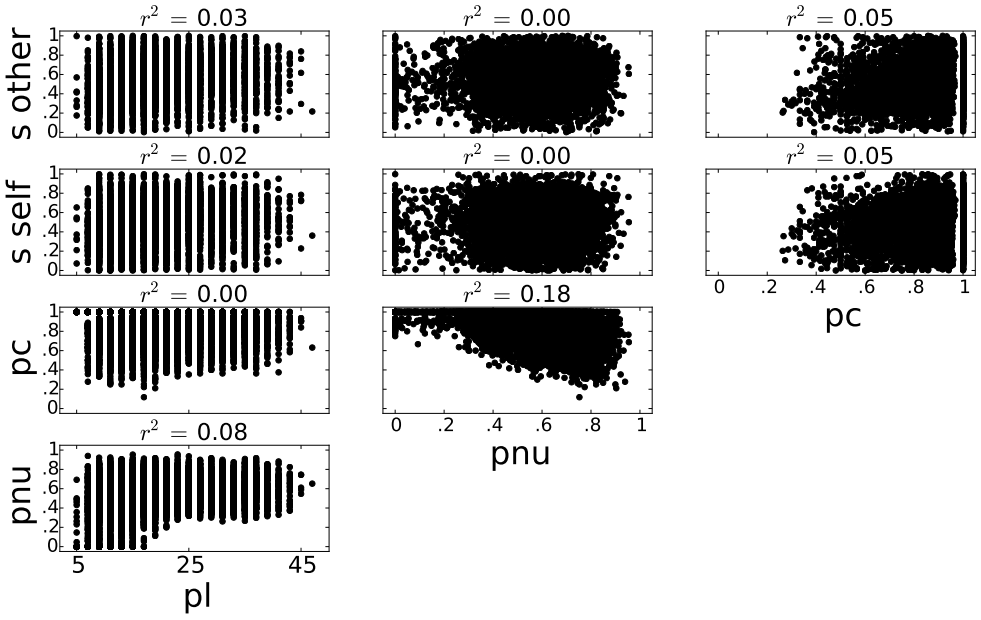


Figure L.1: Correlations between the features of interest. pl: path length, pnu: proportion of non up orientations, pc: proportion of correct key presses, s: skill. Data pooled from all subjects.

Appendix M

T-statistics correction

Permutation methods can be used to protect statistical conclusions about any underlying test. However, not all tests are equally powerful – here because of different numbers of relevant trials. We therefore chose to account for different uncertainties in our measurements of attribution proportions by using weighted measurements for computing the t statistic. We obtained confidence intervals for the differences in proportions that we were interested in, using the continuity correction for the Wald Z estimate, then we weighted differences based on uncertainties associated with these confidence intervals and computed the weighted relevant statistics, thus taking into account the uncertainty associated with each observed difference in proportions.

Specifically, for each subject and each difference in proportion that we are interested in, the estimate for the uncertainty of the proportion difference is obtained as:

$$\sigma = \frac{z * \left[\sqrt{\frac{p_1 * (1-p_1)}{n_1} + \frac{p_2 * (1-p_2)}{n_2}} + \frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \right]}{2}, \text{ where}$$

$z =$ the 95 % percentile from the standard normal distribution

n_1, n_2 are the number of trials in the two conditions we are interested in comparing

p_1, p_2 are the proportions of attributions to the option of interest

in the two conditions we want to compare.

We thus obtained for each subject proportions of attributions in the two conditions - $\left\{p_1^i | i = 1,N\right\}, \left\{p_2^i | i = 1,N\right\}$ - and estimates for the uncertainty associated with the proportion difference for each subject - $\{\sigma_i | i = 1,N\}$; we then computed the weighted repeated measures t-test statistics as follows:

$$t_s = \frac{\overline{\mu}}{\overline{\sigma}}, \text{ where}$$

$$\overline{\mu} = \frac{\Sigma \frac{p_1^i - p_2^i}{\sigma_i^2}}{\Sigma \frac{1}{\sigma_i^2}} \text{ and}$$

$$\overline{\sigma} = \frac{\sqrt{\frac{\Sigma \frac{(p_1^i - p_2^i - \overline{\mu})^2}{\sigma_i^2}}{\Sigma \frac{1}{\sigma_i^2} - 1}}}{\sqrt{N}}$$

We then performed permutation tests (2000 permutations) to compute p-values for t_s.

Appendix N

Outcome effect on attribution

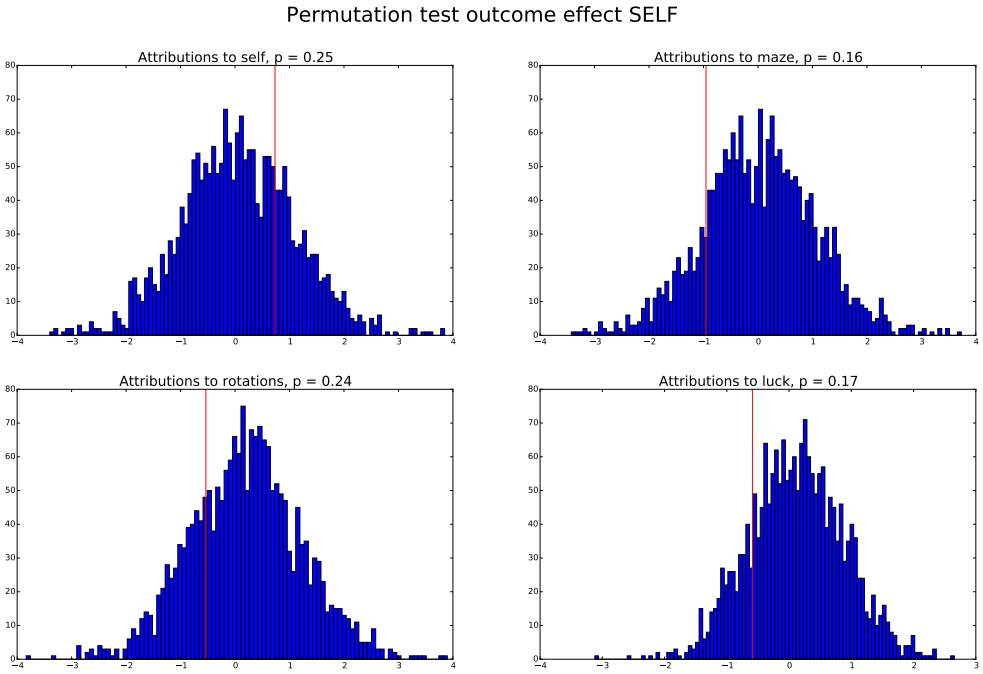


Figure N.1: Effect of outcome on attributions in the "self" condition, results of permutation test. The permutation test was performed as follows: for each iteration (2000), outcome labels for each subject were permuted; the proportion of attributions to each option out of all trials labeled as wins (losses respectively) was computed, along with uncertainty estimates based on small correction Wald Z; the weighted repeated measures t-statistic for the difference between the proportion of attributions after trials labeled as wins and trials labeled as losses was computed across subjects. Histogram shows the resulting statistics values, the red line marks the value of the statistic obtained from the data. The p-value is approximated as the proportion of statistic values beyond the red line.

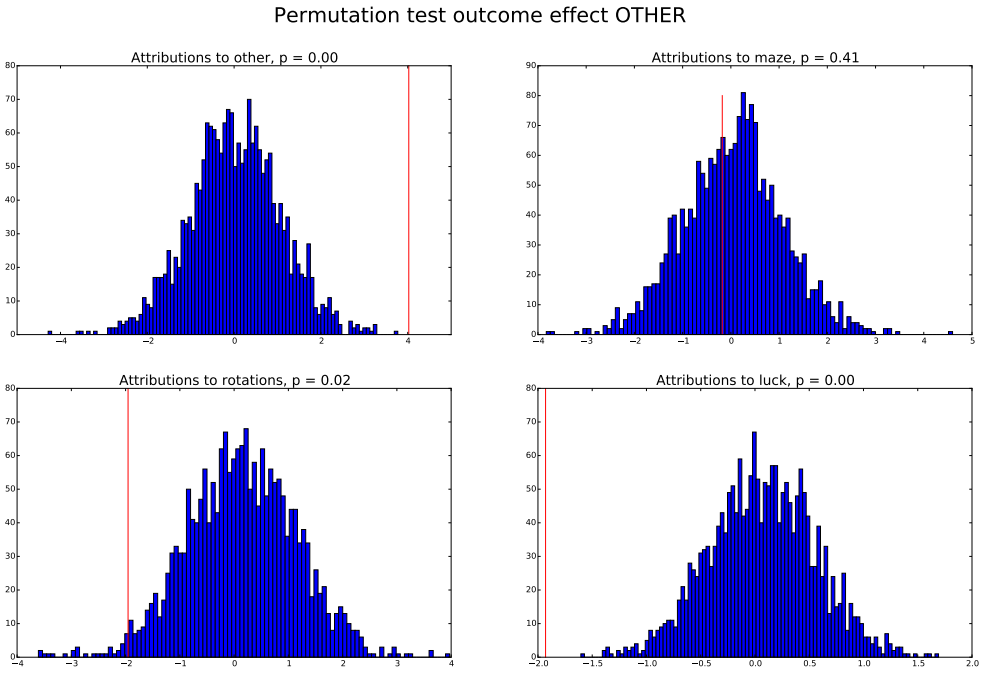


Figure N.2: Effect of outcome on attributions in the "other" condition, results of permutation test. The permutation test was performed as follows: for each iteration (2000), outcome labels for each subject were permuted; the proportion of attributions to each option out of all trials labeled as wins (losses respectively) was computed, along with uncertainty estimates based on small correction Wald Z; the weighted repeated measures t-statistic for the difference between the proportion of attributions after trials labeled as wins and trials labeled as losses was computed across subjects. Histogram shows the resulting statistics values, the red line marks the value of the statistic obtained from the data. The p-value is approximated as the proportion of statistic values beyond the red line.

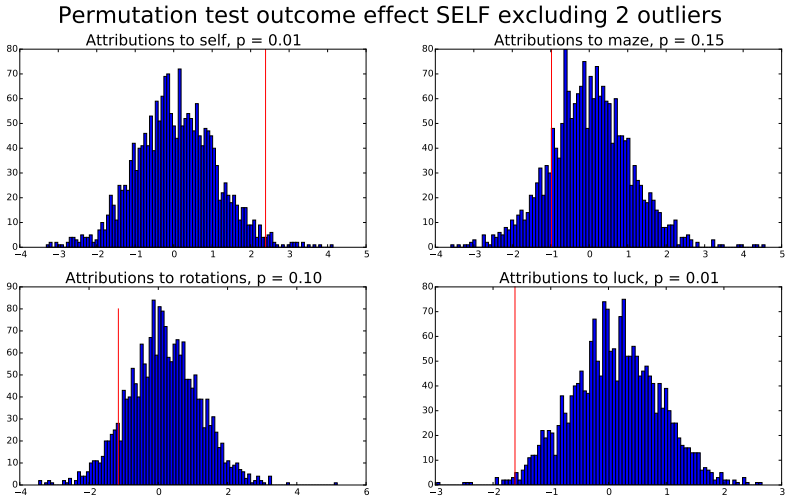


Figure N.3: Effect of outcome on attributions in the "self" condition, excluding the two subjects who made no internal attribution for wins. The permutation test was performed as above.

Appendix O

Outcome and time effect on attribution, permutation tests with 2-way repeated measures ANOVA statistics

The permutation tests were performed on the F-statistics for repeated measures 2-way ANOVA with two fixed within subject factors (outcome and time) and subjects as a random factor. The test was performed for each condition and each attribution option, with the proportion of attributions to the attribution option of interest as the dependent variable.

Let A and B be the factors of interest; a, b, s , the number of levels of factor A, the number of levels of factor B and number of subjects respectively. Let Y_{ijk} be the observation from the i-th subject corresponding to the j-th level of factor A and k-th level of factor B. Overbars denote averages in the corresponding conditions. The F-statistics used were computed as follows Howell (2012):

$$F_A = \frac{MSA}{MSAS}, F_B = \frac{MSB}{MSBS}, F_{AB} = \frac{MSAB}{MSABS}, \text{ where}$$
$$MSA = \frac{SSA}{a-1}, MSAS = \frac{SSAS}{(a-1)(s-1)}$$

$$MSB = \frac{SSB}{b-1}, MSBS = \frac{SSBS}{(b-1)(s-1)}$$
$$MSAB = \frac{SSAB}{(a-1)(b-1)}, MSABS = \frac{SSABS}{(a-1)(b-1)(s-1)}$$
$$SSA = sb\Sigma_j(\overline{Y_{.j.}} - \overline{Y_{...}})^2, SSB = sa\Sigma_k(\overline{Y_{..k}} - \overline{Y_{...}})^2$$
$$SSAS = b\Sigma_{i,j}(\overline{Y_{ij.}} - \overline{Y_{i..}} - \overline{Y_{.j.}} + \overline{Y_{...}})^2$$
$$SSBS = a\Sigma_{i,k}(\overline{Y_{i.k}} - \overline{Y_{i..}} - \overline{Y_{..k}} + \overline{Y_{...}})^2$$
$$SSAB = s\Sigma_{j,k}(\overline{Y_{.jk}} - \overline{Y_{.j.}} - \overline{Y_{..k}} + \overline{Y_{...}})^2$$
$$SSABS = \Sigma_{i,j,k}(Y_{ijk} - \overline{Y_{ij.}} - \overline{Y_{i.k}} - \overline{Y_{.jk}} + \overline{Y_{i..}} + \overline{Y_{.j.}} + \overline{Y_{..k}} - \overline{Y_{...}})^2.$$

Permutation tests corresponding to 2-way ANOVAs with outcome and another factor of interest were performed as follows: for each iteration (5000), we permuted outcome labels and attribution labels for each subject; we computed the proportions of attributions to each option out of all trials labeled as wins (losses respectively) for each level of the other factor of interest; we computed the repeated measures F-statistics for a 2-way ANOVA with outcome and the other factor of interest as fixed factors and subjects as random factors.

In general, the permutation distributions of these statistics turned out to be good approximations of the theoretical F-distributions that the corresponding ANOVA tests would use. In most cases the quality of the approximation is similar to the one illustrated in figure O.1 below; figure O.2 shows the case with the worst match between the permutation distribution and theoretical one.

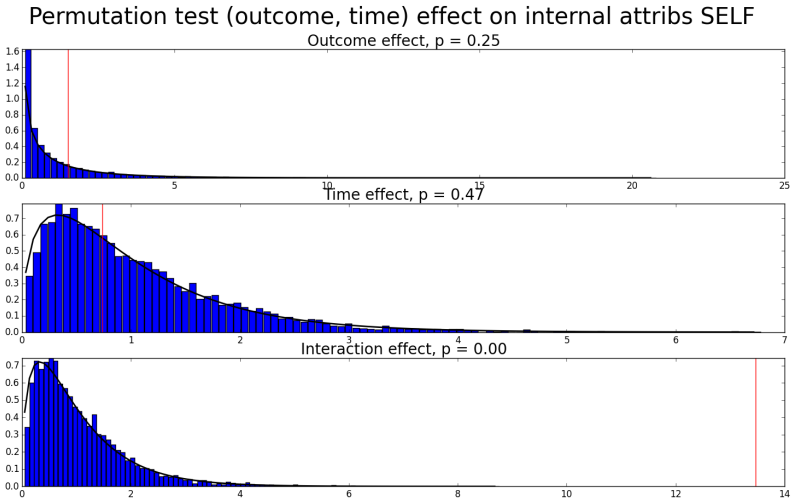


Figure O.1: Effect of outcome and time on attributions to "self", results of permutation test. Histogram shows the distribution of resulting statistics values, in density form; the red line marks the value of the statistic obtained from the data. The p-value is approximated as the proportion of statistic values beyond the value obtained for the unscrambled data. The black curve is the pdf of the F-distribution that the corresponding ANOVA test would use, for comparison with the permutation distribution.

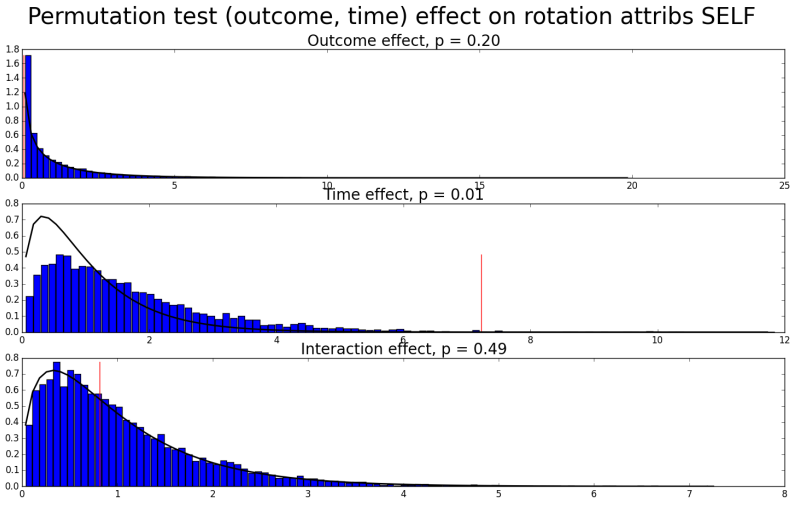


Figure O.2: Effect of outcome and time on attributions to "rotations", in the "self" condition, results of permutation test. The permutation test and p-value computation were performed as described above.

Appendix P

Objective task measures quantization and distributions of outcomes

For the purposes of our model agnostic analyses we discretised continuous factors of interest, such as task and performance features. Ideally, these factors should be independent of outcome, and a fine-tuned staircase adaptation procedure would contribute to achieving this aim. As we did not have previously validated measures of difficulty for our task, the staircase adaptation mechanism that we used turned out to only partially achieve this goal, being rather coarse (see 2). Figures in this appendix show the dependence between objective task features and outcome.

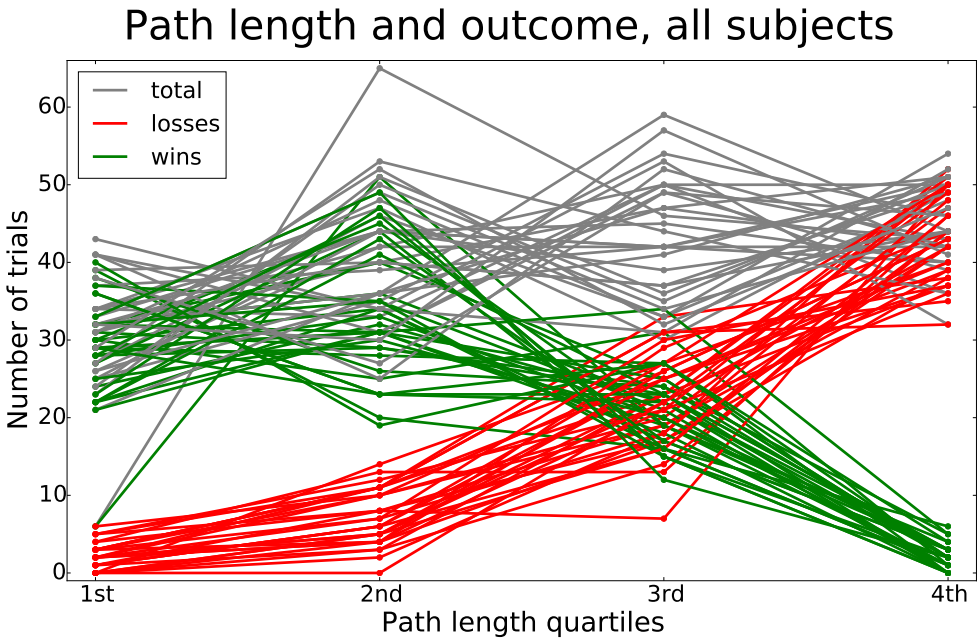


Figure P.1: Path length quantization: total number of trials and number of wins and losses per quantile of path length. Each line represents one subject.

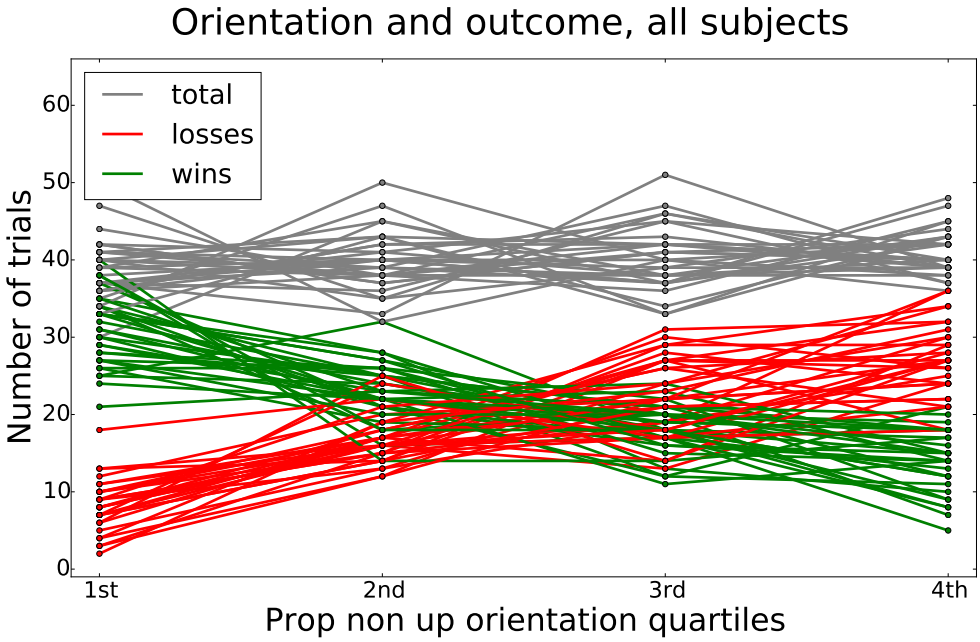


Figure P.2: Prop non up orientations quantization: total number of trials and number of wins and losses per quantile of prop non up orientations. Each line represents one subject.

Appendix Q

Time and skill level quantization

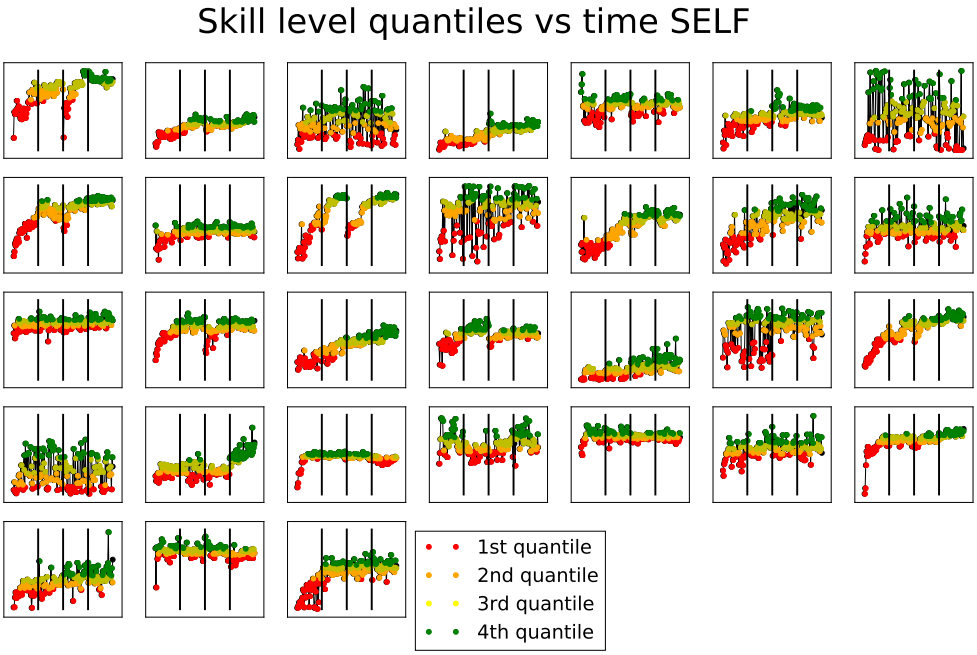


Figure Q.1: Time vs skill quartiles of skill estimates, all subjects 'self' condition: vertical bars mark the time division of trials; dot colors mark skill quantiles, from bottom (red) to top (green).

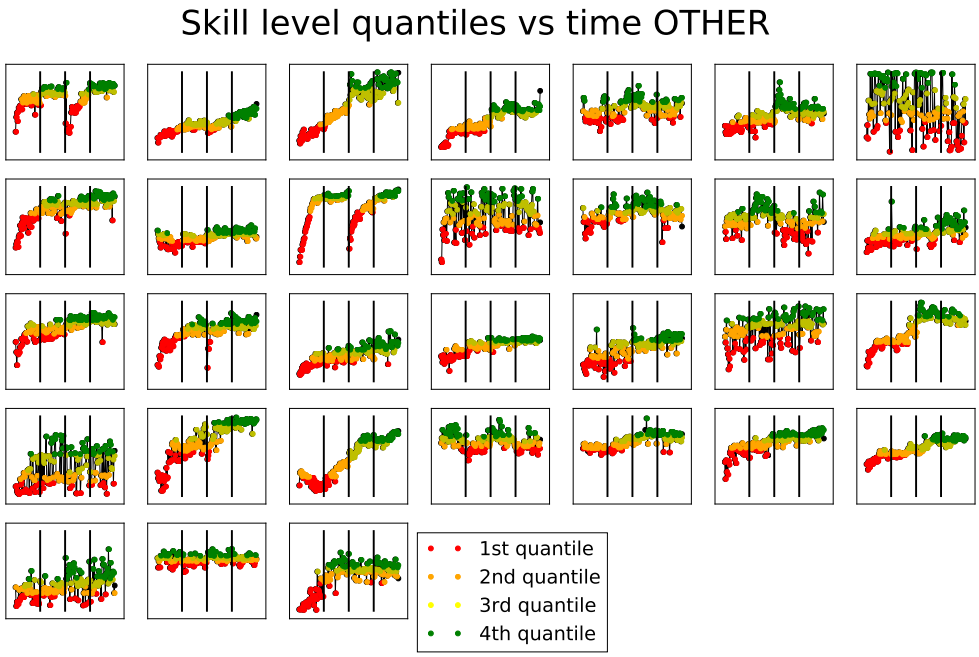


Figure Q.2: Time vs skill quartiles of skill estimates, all subjects 'other' condition: vertical bars mark the time division of trials; dot colors mark skill quantiles, from bottom (red) to top (green).

Appendix R

Models for attribution responses

The models we fit all used linear combinations of features, with parameters acting as weights of the features in computing a score for each possible attribution response. Scores for the different options are passed through a softmax function, to provide response probabilities for the different response options:

$$
\begin{aligned}
s_{t,o} &= \boldsymbol{w}_o \cdot \boldsymbol{f}_t \forall o \in O \\
p_t(o) &= \frac{\exp(s_{t,o})}{\Sigma_{o \in O} \exp(s_{t,o})}, \text{ where} \\
s_{t,o} &= \text{score of response option o on trial t} \\
\boldsymbol{w}_o &= \text{feature weights for option o} \\
\boldsymbol{f}_t &= \text{feature values on trial t} \\
O &= \text{set of available response options} \\
p_t(o) &= \text{probability of choosing option o on trial t}
\end{aligned}
\tag{R.1}
$$

The full set of features used comprised previous skill response (s), length of correct path through the maze (pl), proportion of frames within the trial when the maze orientation was not UP (pnu), proportion of correct key presses (pc), average proportion of pauses in a 2s interval post maze rotation (pp), along with a bias term (b).

Parameters represented weights of the different features for computing scores of the different response options, separately for positive (+) and negative (-) outcomes. For each outcome, weights for whole set of attribution

options were constrained to sum to 0. For the case of binary relabelling of the responses into "internal" and "external", the resulting set of parameters (for the full model):

$$
\begin{aligned}
&\boldsymbol{w}_{I+} = (b_+, \alpha_{s+}, \alpha_{pl+}, \alpha_{pnu+}, \alpha_{pc+}, \alpha_{pp+}) \\
&\qquad \text{for computing the score for internal attributions post win} \\
&\boldsymbol{w}_{I-} = (b_-, \alpha_{s-}, \alpha_{pl-}, \alpha_{pnu-}, \alpha_{pc-}, \alpha_{pp-}) \\
&\qquad \text{for computing the score for internal attributions post loss} \\
&\boldsymbol{w}_{E+} = -\boldsymbol{w}_{I+} \\
&\boldsymbol{w}_{E-} = -\boldsymbol{w}_{I-}
\end{aligned}
\tag{R.2}
$$

When labelling parameters in plots, we use, for brevity, the following convention: if x denotes the feature of interest and o a given outcome, we label α_{xo} as xo (e. g α_{s+} will be labelled as s+). For the case of the 4 response options, the resulting parameter set was:

$$
\begin{aligned}
&\boldsymbol{w}_{I+} = (b_{I+}, \alpha_{sI+}, \alpha_{plI+}, \alpha_{pnuI+}, \alpha_{pcI+}, \alpha_{ppI+}) \\
&\qquad \text{for computing the score for internal attributions post win} \\
&\boldsymbol{w}_{I-} = (b_{I-}, \alpha_{sI-}, \alpha_{plI-}, \alpha_{pnuI-}, \alpha_{pcI-}, \alpha_{ppI-}) \\
&\qquad \text{for computing the score for internal attributions post loss} \\
&\boldsymbol{w}_{M+} = (b_{M+}, \alpha_{sM+}, \alpha_{plM+}, \alpha_{pnuM+}, \alpha_{pcM+}, \alpha_{ppM+}) \\
&\qquad \text{for computing the score for attributions to maze post win} \\
&\boldsymbol{w}_{M-} = (b_{M-}, \alpha_{sM-}, \alpha_{plM-}, \alpha_{pnuM-}, \alpha_{pcM-}, \alpha_{ppM-}) \\
&\qquad \text{for computing the score for attributions to maze post loss} \\
&\boldsymbol{w}_{R+} = (b_{R+}, \alpha_{sR+}, \alpha_{plR+}, \alpha_{pnuR+}, \alpha_{pcR+}, \alpha_{ppR+}) \\
&\qquad \text{for computing the score for attributions to rotations post win} \\
&\boldsymbol{w}_{R-} = (b_{R-}, \alpha_{sR-}, \alpha_{plR-}, \alpha_{pnuR-}, \alpha_{pcR-}, \alpha_{ppR-}) \\
&\qquad \text{for computing the score for attributions to rotations post loss}
\end{aligned}
$$

$$\begin{aligned} \boldsymbol{w}_{L+} &= -(\boldsymbol{w}_{I+} + \boldsymbol{w}_{M+} + \boldsymbol{w}_{R+}) \\ \boldsymbol{w}_{L-} &= -(\boldsymbol{w}_{I-} + \boldsymbol{w}_{M-} + \boldsymbol{w}_{R-}). \end{aligned}$$

When labelling parameters in plots, we use, for brevity, the following convention: if x denotes the feature of interest, A a given attribution response option, and o a given outcome, we label α_{xAo} as xAo (e. g α_{sI+} will be labelled as sI+).

Appendix S

Computing feature effects

For a full description of the attribution models see appendix R. We remind that for the case of the 4 response options, the parameter set of the winning model was[1]:

$$w_{I+} = (b_{I+}, \alpha_{sI+}, \alpha_{plI+}, \alpha_{pnuI+}, \alpha_{pcI+}, \alpha_{ppI+})$$

for computing the score for internal attributions post win

$$w_{I-} = (b_{I-}, \alpha_{sI-}, \alpha_{plI-}, \alpha_{pnuI-}, \alpha_{pcI-}, \alpha_{ppI-})$$

for computing the score for internal attributions post loss

$$w_{M+} = (b_{M+}, \alpha_{sM+}, \alpha_{plM+}, \alpha_{pnuM+}, \alpha_{pcM+}, \alpha_{ppM+})$$

for computing the score for attributions to maze post win

$$w_{M-} = (b_{M-}, \alpha_{sM-}, \alpha_{plM-}, \alpha_{pnuM-}, \alpha_{pcM-}, \alpha_{ppM-})$$

for computing the score for attributions to maze post loss

$$w_{R+} = (b_{R+}, \alpha_{sR+}, \alpha_{plR+}, \alpha_{pnuR+}, \alpha_{pcR+}, \alpha_{ppR+})$$

for computing the score for attributions to rotations post win

$$w_{R-} = (b_{R-}, \alpha_{sR-}, \alpha_{plR-}, \alpha_{pnuR-}, \alpha_{pcR-}, \alpha_{ppR-})$$

for computing the score for attributions to rotations post loss

[1] See glossary for feature definitions and abbreviations

$w_{L+} = -(w_{I+} + w_{M+} + w_{R+})$

for computing the score for attributions to luck post win

.

$w_{L-} = -(w_{I-} + w_{M-} + w_{R-})$

for computing the score for attributions to luck post loss.

Let x denote the feature of interest; we want to compute its contribution to choosing attribution option A having encountered outcome o. We denote this contribution by xAo, e. g sI+ for the contribution of skill to making internal attributions for wins.

$$
\begin{aligned}
xAo &= \frac{1}{T}\Sigma_t \frac{\partial p_t(A)}{\partial x}|_{x=0}, \text{where} \\
T &= \text{total number of trials} \\
p_t(A) &= \text{the probability of choosing attribution option A on trial t} \\
&= \frac{\exp(w_{Ao} \cdot f_t)}{\exp(w_{Io} \cdot f_t) + \exp(w_{Mo} \cdot f_t) + \exp(w_{Ro} \cdot f_t) + \exp(w_{Lo} \cdot f_t)}, \text{where} \\
f_t &= \text{feature values on trial t.}
\end{aligned}
\tag{S.1}
$$

Thus the derivative of $p_t(A)$, seen as a function of x, is evaluated at $x = 0$, and values are then averaged over all trials.

An alternative way of computing the contribution of feature x is to discard all other features, reducing the model to biases and x alone, then computing the derivative of the resulting probability of choosing A after outcome o. We obtained very similar results for the two ways of computing. In the text of the chapter, we presented the ones obtained using the first way of computing xAo.

Bibliography

L. Y. Abramson, M. E. Seligman, and J. D. Teasdale. Learned helplessness in humans: Critique and reformulation. *Journal of Abnormal Psychology*, 87 (1):49–74, 1978. ISSN 0021843X. doi: 10.1037/0021-843X.87.1.49.

L. Y. Abramson, G. I. Metalsky, and L. B. Alloy. Hopelessness depression: A theory-based subtype of depression. *Psychological review*, 96(2):358, 1989.

R. A. Adams, Q. J. Huys, and J. P. Roiser. Computational Psychiatry: Towards a mathematically informed understanding of mental illness. *Journal of Neurology, Neurosurgery and Psychiatry*, 87(1):53–63, 2016. ISSN 1468330X. doi: 10.1136/jnnp-2015-310737.

M. A. Addicott, J. M. Pearson, B. Froeliger, M. L. Platt, and F. Joseph McClernon. Smoking automaticity and tolerance moderate brain activation during explore-exploit behavior. *Psychiatry Research - Neuroimaging*, 224 (3):254–261, 2014. ISSN 18727506. doi: 10.1016/j.pscychresns.2014.10.014. URL `http://dx.doi.org/10.1016/j.pscychresns.2014.10.014`.

M. A. Addicott, J. M. Pearson, M. M. Sweitzer, D. L. Barack, and M. L. Platt. A Primer on Foraging and the Explore/Exploit Trade-Off for Psychiatry Research. *Neuropsychopharmacology*, 42(10):1931–1939, 2017. ISSN 1740634X. doi: 10.1038/npp.2017.108. URL `http://dx.doi.org/10.1038/npp.2017.108`.

H. Akaike. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974. ISSN 15582523. doi: 10.1109/TAC.1974.1100705.

R. Akaishi, N. Kolling, J. W. Brown, and M. Rushworth. Neural mechanisms of credit assignment in a multicue environment. *Journal of Neuroscience*, 36(4):1096–1112, 2016. ISSN 15292401. doi: 10.1523/JNEUROSCI.3159-15.2016.

S. A. Akimoto and D. M. Sanbonmatsu. Differences in self-effacing behavior between European and Japanese Americans: Effect on competence evaluations. *Journal of Cross-Cultural Psychology*, 30(2):159–177, 1999. ISSN 00220221. doi: 10.1177/0022022199030002002.

L. G. Allan, S. Siegel, and S. Hannah. The sad truth about depressive realism. *Quarterly Journal of Experimental Psychology*, 60(3):482–495, 2007. ISSN 17470218. doi: 10.1080/17470210601002686.

L. B. Alloy and L. Y. Abramson. Judgment of contingency in depressed and nondepressed students: Sadder but wiser? *Journal of Experimental Psychology: General*, 108(4):441–485, 1979. ISSN 0096-3445. doi: 10.1037//0096-3445.108.4.441.

L. B. Alloy and L. Y. Abramson. Learned helplessness, depression, and the illusion of control. *Journal of Personality and Social Psychology*, 42(6):1114–1126, 1982. ISSN 0022-3514. doi: 10.1037//0022-3514.42.6.1114.

L. B. Alloy, L. Y. Abramson, and D. Viscusi. Induced mood and the illusion of control. *Journal of Personality and Social Psychology*, 41(6):1129–1140, 1981. ISSN 00223514. doi: 10.1037/0022-3514.41.6.1129.

L. B. Alloy, L. Y. Abramson, W. G. Whitehouse, M. E. Hogan, N. A. Tashman, D. L. Steinberg, D. T. Rose, and P. Donovan. Depressogenic cognitive styles: Predictive validity, information processing and personality

characteristics, and developmental origins. *Behaviour Research and Therapy*, 37(6):503–531, 1999. ISSN 00057967. doi: 10.1016/S0005-7967(98)00157-0.

L. B. Alloy, W. G. Whitehouse, M. S. Robinson, J. B. Lapkin, L. Y. Abramson, M. E. Hogan, D. T. Rose, and R. S. Kim. The Temple-Wisconsin Cognitive Vulnerability to Depression Project: Lifetime History of Axis I Psychopathology in Individuals at High and Low Cognitive Risk for Depression. *Journal of Abnormal Psychology*, 109(3):403–418, 2000. doi: 10.1037//0021-843X.109.3,403.

W. F. Asaad, P. M. Lauro, J. A. Perge, and E. N. Eskandar. Prefrontal neurons encode a solution to the credit-assignment problem. *Journal of Neuroscience*, 37(29):6995–7007, 2017. ISSN 15292401. doi: 10.1523/JNEUROSCI.3311-16.2017.

A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, and J. Erbaugh. An inventory for measuring depression. *Archives of general psychiatry*, 4(6):561–571, 1961.

G. M. Becker, M. H. DeGroot, and J. Marschak. Measuring utility by a single-response sequential method. *Behavioral science*, 9(3):226–232, 1964.

T. E. Behrens, M. W. Woolrich, M. E. Walton, and M. F. Rushworth. Learning the value of information in an uncertain world. *Nature neuroscience*, 10(9):1214–1221, 2007.

T. E. Behrens, L. T. Hunt, and M. F. Rushworth. The computation of social behavior. *Science*, 324(5931):1160–1164, 2009. ISSN 00368075. doi: 10.1126/science.1169694.

V. A. Benassi and H. I. Mahler. Contingency Judgments By Depressed College Students. Sadder But Not Always Wiser. *Journal of Personality and Social Psychology*, 49(5):1323–1329, 1985. ISSN 00223514. doi: 10.1037/0022-3514.49.5.1323.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

R. P. Bentall. *Madness explained: Psychosis and human nature*. Penguin UK, 2003.

R. P. Bentall and S. Kaney. Attributional lability in depression and paranoia. *British Journal of Clinical Psychology*, 44(4):475–488, 2005.

R. P. Bentall and M. Thompson. Emotional Stroop performance and the manic defence. *British Journal of Clinical Psychology*, 29(2):235–237, 1990. ISSN 20448260. doi: 10.1111/j.2044-8260.1990.tb00877.x.

R. P. Bentall, R. Corcoran, R. Howard, N. Blackwood, and P. Kinderman. Persecutory delusions: a review and theoretical integration. *Clinical psychology review*, 21(8):1143–1192, 2001.

E. Berglund, P. Lytsy, and R. Westerling. The influence of locus of control on self-rated health in context of chronic disease: A structural equation modeling approach in a cross sectional study. *BMC Public Health*, 14(1): 1–9, 2014. ISSN 14712458. doi: 10.1186/1471-2458-14-492.

C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006.

B. Blaine and J. Crocker. Self-esteem and self-serving biases in reactions to positive and negative events: An integrative review. In *Self-esteem*, pages 55–85. Springer, 1993.

N. J. Blanco, A. R. Otto, W. T. Maddox, C. G. Beevers, and B. C. Love. The influence of depression symptoms on exploratory decision-making. *Cognition*, 129(3):563–568, 2013. ISSN 00100277. doi: 10.1016/j.cognition.2013.08.018. URL `http://dx.doi.org/10.1016/j.cognition.2013.08.018`.

G. Bohner, H. Bless, N. Schwarz, and F. Strack. What triggers causal attributions? The impact of valence and subjective probability. *European Journal of Social Psychology*, 18(4):335–345, 1988. ISSN 10990992. doi: 10.1002/ejsp.2420180404.

C. Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62, 1936.

J. F. Brosschot, W. A. Gebhardt, and G. L. Godaert. Internal, powerful others and chance locus of control: Relationships with personality, coping, stress and health. *Personality and individual Differences*, 16(6):839–852, 1994.

M. C. Butler and R. G. Burr. Utility of a multidimensional locus of control scale in predicting health and job-related outcomes in military environments. *Psychological Reports*, 47(3 Pt 1):719–728, 1980. ISSN 00332941. doi: 10.2466/pr0.1980.47.3.719.

W. K. Campbell and C. Sedikides. Self-threat magnifies the self-serving bias: A meta-analytic integration. *Review of general Psychology*, 3(1):23–43, 1999.

R. C. Carson, S. D. Hollon, and R. C. Shelton. Depressive realism and clinical depression. *Behaviour Research and Therapy*, 48(4):257–265, 2010. ISSN 00057967. doi: 10.1016/j.brat.2009.11.011. URL `http://dx.doi.org/10.1016/j.brat.2009.11.011`.

R. D. Cazé and M. A. Van Der Meer. Adaptive properties of differential learning rates for positive and negative outcomes. *Biological Cybernetics*, 107(6):711–719, 2013. ISSN 03401200. doi: 10.1007/s00422-013-0571-5.

M. Cella, S. Dymond, and A. Cooper. Impaired flexible decision-making in major depressive disorder. *Journal of Affective Disorders*, 124(1-2):207–210, 2010. ISSN 01650327. doi: 10.1016/j.jad.2009.11.013. URL `http://dx.doi.org/10.1016/j.jad.2009.11.013`.

R. Cools, O. J. Robinson, and B. Sahakian. Acute tryptophan depletion in healthy volunteers enhances punishment prediction but does not affect reward prediction. *Neuropsychopharmacology*, 33(9):2291–2299, 2008. ISSN 0893133X. doi: 10.1038/sj.npp.1301598.

S. M. Cox, M. J. Frank, K. Larcher, L. K. Fellows, C. A. Clark, M. Leyton, and A. Dagher. Striatal D1 and D2 signaling differentially predict learning from positive and negative outcomes. *NeuroImage*, 109:95–101, 2015. ISSN 10959572. doi: 10.1016/j.neuroimage.2014.12.070. URL `http://dx.doi.org/10.1016/j.neuroimage.2014.12.070`.

J. Coyne and I. H. Gotlib. Coyne_1983_Cognition_Depression. *Psychological Bulletin*, 94(3):472–505, 1983.

M. J. Crockett, Z. Kurth-Nelson, J. Z. Siegel, P. Dayan, and R. J. Dolan. Harm to others outweighs harm to self in moral decision making. *Proceedings of the National Academy of Sciences of the United States of America*, 111(48): 17320–17325, 2014. ISSN 10916490. doi: 10.1073/pnas.1408988111.

J. Daunizeau, H. E. den Ouden, M. Pessiglione, S. J. Kiebel, K. J. Friston, and K. E. Stephan. Observing the observer (II): Deciding when to decide. *PLoS ONE*, 5(12), 2010a. ISSN 19326203. doi: 10.1371/journal.pone.0015555.

J. Daunizeau, H. E. den Ouden, M. Pessiglione, S. J. Kiebel, K. E. Stephan, and K. J. Friston. Observing the observer (I): Meta-bayesian models of learning and decision-making. *PLoS ONE*, 5(12), 2010b. ISSN 19326203. doi: 10.1371/journal.pone.0015554.

N. D. Daw and K. Doya. The computational neurobiology of learning and reward. *Current Opinion in Neurobiology*, 16(2):199–204, 2006. ISSN 09594388. doi: 10.1016/j.conb.2006.03.006.

N. D. Daw, S. J. Gershman, B. Seymour, P. Dayan, and R. J. Dolan. Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6):1204–1215, 2011.

P. Dayan and L. F. Abbott. *Theoretical neuroscience: computational and mathematical modeling of neural systems*. Computational Neuroscience Series, 2001.

P. Dayan and N. D. Daw. Decision theory, reinforcement learning, and the brain. *Cognitive, Affective and Behavioral Neuroscience*, 8(4):429–453, 2008. ISSN 15307026. doi: 10.3758/CABN.8.4.429.

P. Dayan and Y. Niv. Reinforcement learning: The Good, The Bad and The Ugly. *Current Opinion in Neurobiology*, 18(2):185–196, 2008. ISSN 09594388. doi: 10.1016/j.conb.2008.08.003.

K. Deaux and E. Farris. Attributing causes for one's own performance: The effects of sex, norms, and outcome. *Journal of Research in Personality*, 11 (1):59–72, 1977. ISSN 10957251. doi: 10.1016/0092-6566(77)90029-0.

L. Devenport. Superstitious bar pressing in hippocampal and septal rats. *Science*, 205(4407):721–723, 1979.

A. Dickinson, A. Watt, and W. J. Griffiths. Free-operant Acquisition with Delayed Reinforcement. *The Quarterly Journal of Experimental Psychology Section B*, 45(3):241–258, 1992. ISSN 14641321. doi: 10.1080/14640749208401019.

D. Dunning, A. Leuenberger, and D. A. Sherman. A New Look at Motivated Inference: Are Self-Serving Theories of Success a Product of Motivational Forces? *Journal of Personality and Social Psychology*, 69(1):58–68, 1995. ISSN 00223514. doi: 10.1037/0022-3514.69.1.58.

E. Eldar and Y. Niv. Interaction between emotional state and learning underlies mood instability. *Nature communications*, 6(1):1–10, 2015.

F. D. Fincham. Outcome Valence and Situational Constraints in the Responsibility Attributions of Children and Adults. *Social Cognition*, 3(2):218–233, 1985. ISSN 0278-016X. doi: 10.1521/soco.1985.3.2.218.

T. H. FitzGerald, P. Schwartenbeck, M. Moutoussis, R. J. Dolan, and K. Friston. Active inference, evidence accumulation, and the urn task. *Neural computation*, 27(2):306–328, 2015.

J. P. Forgas, G. H. Bower, and S. J. Moylan. Praise or blame? Affective influences on attributions for achievement. *Journal of Personality and Social Psychology*, 59(4):809–819, 1990. ISSN 0022-3514. doi: 10.1037//0022-3514.59.4.809.

M. J. Frank, A. A. Moustafa, H. M. Haughey, T. Curran, and K. E. Hutchison. Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences of the United States of America*, 104(41):16311–16316, 2007. ISSN 00278424. doi: 10.1073/pnas.0706111104.

C. C. French, A. Richards, and E. J. Scholfield. Hypomania, anxiety and the emotional Stroop. *British Journal of Clinical Psychology*, 35(4):617–626, 1996. ISSN 01446657. doi: 10.1111/j.2044-8260.1996.tb01217.x.

J. Garber and S. D. Hollon. Universal versus personal helplessness in depression: Belief in uncontrollability or incompetence? *Journal of Abnormal Psychology*, 89(1):56–66, 1980. ISSN 0021843X. doi: 10.1037/0021-843X.89.1.56.

J. Garrison, B. Erdeniz, and J. Done. Prediction error in reinforcement learning: A meta-analysis of neuroimaging studies. *Neuroscience and Biobehavioral Reviews*, 37(7):1297–1310, 2013. ISSN 01497634. doi: 10.1016/j.neubiorev.2013.03.023. URL `http://dx.doi.org/10.1016/j.neubiorev.2013.03.023`.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. CRC press, 2013.

A. Gelman, J. Hwang, and A. Vehtari. Understanding predictive information

criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016, 2014. ISSN 15731375. doi: 10.1007/s11222-013-9416-2.

L. Giuntoli, I. Marchetti, A. Panzeri, A. Spoto, G. Vidotto, and C. Caudek. Measuring cognitive vulnerability to depression: Further evidence on the factorial and predictive validity of negative cognitive style. *Journal of Behavior Therapy and Experimental Psychiatry*, 65(October 2018):101479, 2019. ISSN 18737943. doi: 10.1016/j.jbtep.2019.04.005. URL `https://doi.org/10.1016/j.jbtep.2019.04.005`.

P. W. Glimcher. Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, 108(Supplement 3):15647–15654, 2011.

S. Golin, F. Terrell, and B. Johnson. Depression and the illusion of control. *Journal of Abnormal Psychology*, 86(4):440–442, 1977. ISSN 0021843X. doi: 10.1037/0021-843X.86.4.440.

J. J. Goodnow. Determinants of choice-distribution in two-choice situations. *The American Journal of Psychology*, 68(1):106–116, 1955.

I. H. Gotlib. Self-reinforcement and recall: Differential deficits in depressed and nondepressed psychiatric inpatients. *Journal of Abnormal Psychology*, 90(6):521–530, 1981. ISSN 0021843X. doi: 10.1037/0021-843X.90.6.521.

I. H. Gotlib and C. D. McCann. Construct accessibility and depression: An examination of cognitive and affective factors. *Journal of Personality and Social Psychology*, 47(2):427–439, 1984. ISSN 00223514. doi: 10.1037/0022-3514.47.2.427.

G. J. Haeffel and I. Vargas. Resilience to depressive symptoms: The buffering effects of enhancing cognitive style and positive life events. *Journal of behavior therapy and experimental psychiatry*, 42(1):13–18, 2011.

G. J. Haeffel, B. E. Gibb, G. I. Metalsky, L. B. Alloy, L. Y. Abramson, B. L. Hankin, T. E. Joiner, and J. D. Swendsen. Measuring cognitive vulnerability to depression: Development and validation of the cognitive style questionnaire. *Clinical Psychology Review*, 28(5):824–836, 2008. ISSN 02727358. doi: 10.1016/j.cpr.2007.12.001.

T. U. Hauser, G. J. Will, M. Dubois, and R. J. Dolan. Annual Research Review: Developmental computational psychiatry. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 60(4):412–426, 2019. ISSN 14697610. doi: 10.1111/jcpp.12964.

F. Heider. *The psychology of interpersonal relations*. Psychology Press, 1982.

S. J. Heine and T. Hamamura. In search of east Asian self-enhancement. *Personality and Social Psychology Review*, 11(1):4–27, 2007. ISSN 10888683. doi: 10.1177/1088868306294587.

R. J. Herrnstein. Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the experimental analysis of behavior*, 4(3):267, 1961.

A. K. Hewitt, D. R. Foxcroft, and J. MacDonald. Multitrait-multimethod confirmatory factor analysis of the attributional style questionnaire. *Personality and Individual Differences*, 37(7):1483–1491, 2004. ISSN 01918869. doi: 10.1016/j.paid.2004.02.005.

E. T. Higgins. Self-Discrepancy: A Theory Relating Self and Affect. *Psychological Review*, 94(3):319–340, 1987. ISSN 0033295X. doi: 10.1037/0033-295X.94.3.319.

N. Higgins, B. D. Zumbo, and J. L. Hay. Construct validity of attributional style: Modeling context-dependent item sets in the attributional style questionnaire. *Educational and Psychological Measurement*, 59(5):804–820, 1999.

D. S. Hiroto. Locus of control and learned helplessness. *Journal of Experimental Psychology*, 102(2):187–193, 1974. ISSN 00221015. doi: 10.1037/h0035910.

D. C. Howell. *Statistical methods for psychology*. Cengage Learning, 2012.

Q. J. Huys, N. Eshel, E. O'Nions, L. Sheridan, P. Dayan, and J. P. Roiser. Bonsai trees in your head: How the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Computational Biology*, 8(3), 2012. ISSN 1553734X. doi: 10.1371/journal.pcbi.1002410.

Q. J. Huys, M. Guitart-Masip, R. J. Dolan, and P. Dayan. Decision-Theoretic Psychiatry. *Clinical Psychological Science*, 3(3):400–421, 2015. ISSN 21677034. doi: 10.1177/2167702614562040.

G. J. Hyman, R. Stanley, and G. D. Burrows. The relationship between three multidimensional locus of control scales. *Educational and Psychological Measurement*, 51(2):403–412, 1991.

K. Iigaya, A. Jolivald, W. Jitkrittum, I. D. Gilchrist, P. Dayan, E. Paul, and M. Mendl. Cognitive bias in ambiguity judgements: using computational models to dissect the effects of mild mood manipulation in humans. *PloS one*, 11(11):e0165840, 2016.

E. M. Izhikevich. Solving the distal reward problem through linkage of STDP and dopamine signaling. *Cerebral Cortex*, 17(10):2443–2452, 2007. ISSN 10473211. doi: 10.1093/cercor/bhl152.

P. L. Jackson, E. Brunet, A. N. Meltzoff, and J. Decety. Empathy examined through the neural mechanisms involved in imagining how I feel versus how you feel pain. *Neuropsychologia*, 44(5):752–761, 2006. ISSN 00283932. doi: 10.1016/j.neuropsychologia.2005.07.015.

G. Jocham, K. H. Brodersen, A. O. Constantinescu, M. C. Kahn, A. M. Ianni,

M. E. Walton, M. F. Rushworth, and T. E. Behrens. Reward-guided learning with and without causal attribution. *Neuron*, 90(1):177–190, 2016.

J. Johnson, M. Panagioti, J. Bass, L. Ramsey, and R. Harrison. Resilience to emotional distress in response to failure, error or mistakes: A systematic review. *Clinical psychology review*, 52:19–42, 2017.

E. E. Jones and R. E. Nisbett. The actor and the observer: Divergent perceptions of the causes of behavior. In *Preparation of this paper grew out of a workshop on attribution theory held at University of California, Los Angeles, Aug 1969*. Lawrence Erlbaum Associates, Inc, 1987.

L. J. Kamin. Trace conditioning of the conditioned emotional response. *Journal of Comparative and Physiological Psychology*, 54(2):149–153, 1961. ISSN 00219940. doi: 10.1037/h0045611.

S. Kaney and R. P. Bentall. Persecutory delusions and the self-serving bias: Evidence from a contingency judgment task. *Journal of Nervous and Mental Disease*, 1992.

H. H. Kelley. Attribution theory in social psychology. In *Nebraska symposium on motivation*. University of Nebraska Press, 1967.

H. H. Kelley and J. L. Michela. Attribution Theory and Research. *Annual Review of Psychology*, 31(1):457–501, 1980. ISSN 0066-4308. doi: 10.1146/annurev.ps.31.020180.002325.

P. Kinderman and R. P. Bentall. A new measure of causal locus: the internal, personal and situational attributions questionnaire. *Personality and Individual differences*, 20(2):261–264, 1996.

E. M. Kleiman, R. T. Liu, and J. H. Riskind. Enhancing attributional style as a resiliency factor in depressogenic stress generation. *Anxiety, Stress & Coping*, 26(4):467–474, 2013.

A. H. Klopf. *Brain function and adaptive systems: a heterostatic theory*. Number 133. Air Force Cambridge Research Laboratories, Air Force Systems Command, United ?, 1972.

A. H. Klopf. *The hedonistic neuron: a theory of memory, learning, and intelligence*. Toxicology-Sci, 1982.

D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

J. F. Kominsky, J. Phillips, T. Gerstenberg, D. Lagnado, and J. Knobe. Causal superseding. *Cognition*, 137:196–209, 2015. ISSN 18737838. doi: 10.1016/j.cognition.2015.01.013. URL `http://dx.doi.org/10.1016/j.cognition.2015.01.013`.

C. K. Kovach, N. D. Daw, D. Rudrauf, D. Trane, J. P. O'Doherty, and R. Adolphs. Anterior prefrontal cortex contributes to action selection through tracking of recent reward trends. *Journal of Neuroscience*, 32(25): 8434–8442, 2012. ISSN 02706474. doi: 10.1523/JNEUROSCI.5468-11.2012.

N. A. Kuiper and T. B. Rogers. Encoding of personal information: Self-other differences. *Journal of Personality and Social Psychology*, 37(4):499–514, 1979. ISSN 00223514. doi: 10.1037/0022-3514.37.4.499.

D. A. Lagnado and S. Channon. Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, 108(3):754–770, 2008. ISSN 00100277. doi: 10.1016/j.cognition.2008.06.009.

B. Lau and P. W. Glimcher. Value Representations in the Primate Striatum during Matching Behavior. *Neuron*, 58(3):451–463, 2008. ISSN 08966273. doi: 10.1016/j.neuron.2008.02.021.

R. C. Leaf. Avoidance response evocation as a function of prior discriminative

fear conditioning under curare. *Journal of Comparative and Physiological Psychology*, 58(3):446, 1964.

M. P. Lehmann, H. A. Xu, V. Liakoni, M. H. Herzog, W. Gerstner, and K. Preuschoff. One-shot learning and behavioral eligibility traces in sequential decision making. *eLife*, 8(200020):1–32, 2019. ISSN 2050084X. doi: 10.7554/eLife.47463.

H. Levenson. Reliability and validity of the i, p, and c scales-a multidimensional view of locus of control. 1973.

H. Levenson. Activism and Powerful Others: Distinctions Within the Concept of Internal-External Control. *Journal of Personality Assessment*, 38(4): 377–383, 1974. ISSN 15327752. doi: 10.1080/00223891.1974.10119988.

R. T. Liu, E. M. Kleiman, B. A. Nestor, and S. M. Cheek. The hopelessness theory of depression: A quarter-century in review. *Clinical Psychology: Science and Practice*, 22(4):345–365, 2015.

W. C. Lobitz and R. D. Post. Parameters of self-reinforcement and depression. *Journal of Abnormal Psychology*, 88(1):33–41, 1979. ISSN 0021843X. doi: 10.1037/0021-843X.88.1.33.

H. M. Lyon, M. Startup, and R. P. Bentall. Social cognition and the manic defense: Attributions, selective attention, and self-schema in bipolar affective disorder. *Journal of Abnormal Psychology*, 108(2):273, 1999.

B. A. Lyons, A. M. McKay, and J. Reifler. High-status lobbyists are most likely to overrate their success. *Nature Human Behaviour*, 4(2):153–159, 2020.

D. J. MacKay and D. J. Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

S. F. Maier and M. E. Seligman. Learned helplessness: Theory and evidence. *Journal of Experimental Psychology: General*, 105(1):3–46, 1976. ISSN 0096-3445. doi: 10.1037//0096-3445.105.1.3.

S. F. Maier and M. E. Seligman. Learned helplessness at fifty: Insights from neuroscience. *Psychological review*, 123(4):349, 2016.

B. F. Malle. The actor-observer asymmetry in attribution: A (surprising) meta-analysis. *Psychological Bulletin*, 132(6):895–919, 2006. ISSN 00332909. doi: 10.1037/0033-2909.132.6.895.

F. Mancinelli, J. Roiser, and P. Dayan. Subjective beliefs in, out, and about control: A quantitative analysis. *bioRxiv*, 2020.

D. J. Martin, L. Y. Abramson, and L. B. Alloy. Illusion of control for self and others in depressed and nondepressed college students. *Journal of Personality and Social Psychology*, 46(1):125–136, 1984. ISSN 00223514. doi: 10.1037/0022-3514.46.1.125.

J. Martín-Albo, J. L. Núñez, J. G. Navarro, and F. Grijalvo. The Rosenberg self-esteem scale: Translation and validation in university students. *Spanish Journal of Psychology*, 10(2):458–467, 2007. ISSN 19882904. doi: 10.1017/S1138741600006727.

C. Mathys, J. Daunizeau, K. J. Friston, and K. E. Stephan. A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, 5(MAY):9, 2011. ISSN 16625161. doi: 10.3389/fnhum.2011.00039.

S. D. McDougle, M. J. Boggess, M. J. Crossley, D. Parvin, R. B. Ivry, and J. A. Taylor. Credit assignment in movement-dependent reinforcement learning. *Proceedings of the National Academy of Sciences of the United States of America*, 113(24):6797–6802, 2016. ISSN 10916490. doi: 10.1073/pnas.1523669113.

S. D. McDougle, P. A. Butcher, D. E. Parvin, F. Mushtaq, Y. Niv, R. B. Ivry, and J. A. Taylor. Neural signatures of prediction errors in a decision-making task are modulated by action execution failures. *Current Biology*, 29(10):1606–1613, 2019.

E. Meins, S. McCarthy-Jones, C. Fernyhough, G. Lewis, R. P. Bentall, and L. B. Alloy. Assessing negative cognitive style: Development and validation of a Short-Form version of the Cognitive Style Questionnaire. *Personality and Individual Differences*, 52(5):581–585, 2012. ISSN 01918869. doi: 10.1016/j.paid.2011.11.026. URL `http://dx.doi.org/10.1016/j.paid.2011.11.026`.

G. I. Metalsky, L. J. Halberstadt, and L. Y. Abramson. Vulnerability to Depressive Mood Reactions: Toward a More Powerful Test of the Diathesis-Stress and Causal Mediation Components of the Reformulated Theory of Depression. *Journal of Personality and Social Psychology*, 52(2):386–393, 1987. ISSN 00223514. doi: 10.1037/0022-3514.52.2.386.

A. H. Mezulis, L. Y. Abramson, J. S. Hyde, and B. L. Hankin. Is there a universal positivity bias in attributions? A meta-analytic review of individual, developmental, and cultural differences in the self-serving attributional bias. *Psychological Bulletin*, 130(5):711–747, 2004. ISSN 00332909. doi: 10.1037/0033-2909.130.5.711.

A. G. Miller. Actor and observer perceptions of the learning of a task. *Journal of Experimental Social Psychology*, 11(2):95–111, 1975. ISSN 10960465. doi: 10.1016/S0022-1031(75)80014-X.

D. T. Miller and M. Ross. Self-serving biases in the attribution of causality: Fact or fiction? *Psychological Bulletin*, 82(2):213–225, 1975. ISSN 00332909. doi: 10.1037/h0076486.

W. R. Miller and M. E. Seligman. Depression and the perception of reinforcement. *Journal of Abnormal Psychology*, 82(1):62, 1973.

W. R. Miller and M. E. Seligman. Depression and learned helplessness in man. *Journal of Abnormal Psychology*, 84(3):228–238, 1975. ISSN 0021843X. doi: 10.1037/h0076720.

W. R. Miller and M. E. Seligman. Learned helplessness, depression and the perception of reinforcement. *Behaviour Research and Therapy*, 14(1):7–17, 1976. ISSN 00057967. doi: 10.1016/0005-7967(76)90039-5.

M. Minsky. Steps Toward Artificial Intelligence. *Proceedings of the IRE*, 49 (1):8–30, 1961. ISSN 00968390. doi: 10.1109/JRPROC.1961.287775.

P. R. Montague, R. J. Dolan, K. J. Friston, and P. Dayan. Computational psychiatry. *Trends in Cognitive Sciences*, 16(1):72–80, 2012. ISSN 13646613. doi: 10.1016/j.tics.2011.11.018.

M. T. Moore and D. M. Fresco. Depressive realism: A meta-analytic review. *Clinical Psychology Review*, 32(6):496–509, 2012. ISSN 02727358. doi: 10.1016/j.cpr.2012.05.004. URL `http://dx.doi.org/10.1016/j.cpr.2012.05.004`.

P. Moreira, J. M. Vaz, D. Stevanovic, O. Atilola, K. Dodig-Ćurković, T. Franic, A. Djoric, N. Davidovic, M. Avicenna, I. Multazam Noor, M. L. A. Campos, A. Ribas, D. Stupar, A. Deljkovic, L. Nussbaum, A. Thabet, D. Ubalde, P. Petrov, P. Vostanis, R. Knez, Y. P. S. Balhara, N. Ana, M. Paulo, L. A. Monteiro, O. Olanrewaju, and L. Bolanle. Locus of control, negative live events and psychopathological symptoms in collectivist adolescents. *Personality and Individual Differences*, 154(October 2019), 2020. ISSN 01918869. doi: 10.1016/j.paid.2019.109601.

G. Morelli, H. Krotinger, and S. Moore. Neuroticism and levenson's locus of control scale. *Psychological Reports*, 44(1):153–154, 1979.

L. S. Morris, K. Baek, P. Kundu, N. A. Harrison, M. J. Frank, and V. Voon. Biases in the Explore-Exploit Tradeoff in Addictions: The Role of Avoidance

of Uncertainty. *Neuropsychopharmacology*, 41(4):940–948, 2016. ISSN 1740634X. doi: 10.1038/npp.2015.208. URL http://dx.doi.org/10.1038/npp.2015.208.

E. B. Morse and W. N. Runquist. Probability-Matching with an Unscheduled Random Sequence. *The American Journal of Psychology*, 73(4):603–607, 1960. URL http://www.jstor.com/stable/1419951.

L. Müller-Pinzler, N. Czekalla, A. V. Mayer, D. S. Stolz, V. Gazzola, C. Keysers, F. M. Paulus, and S. Krach. Negativity-bias in forming beliefs about own abilities. *Scientific Reports*, 9(1):1–15, 2019. ISSN 20452322. doi: 10.1038/s41598-019-50821-w.

F. Mushtaq, S. McDougle, M. Craddock, D. Parvin, J. Brookes, A. Schaefer, M. Mon-Williams, J. Taylor, and R. Ivry. Distinct Processing of Selection and Execution Errors in Neural Signatures of Outcome Monitoring. *bioRxiv*, page 853317, 2019. doi: 10.1101/853317.

D. J. Needles and L. Y. Abramson. Positive Life Events, Attributional Style, and Hopefulness: Testing a Model of Recovery From Depression. *Journal of Abnormal Psychology*, 99(2):156–165, 1990. ISSN 0021843X. doi: 10.1037//0021-843x.99.2.156.

R. E. Nisbett, C. Caputo, P. Legant, and J. Marecek. Behavior as seen by the actor and as seen by the observer. *Journal of personality and Social Psychology*, 27(2):154, 1973.

Y. Niv, J. A. Edlund, P. Dayan, and J. P. O'Doherty. Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience*, 32(2):551–562, 2012. ISSN 02706474. doi: 10.1523/JNEUROSCI.5498-10.2012.

M. P. Noonan, B. K. Chau, M. F. Rushworth, and L. K. Fellows. Contrasting effects of medial and lateral orbitofrontal cortex lesions on credit assignment and decision-making in humans. *Journal of Neuroscience*, 37(29):

7023–7035, 2017. ISSN 15292401. doi: 10.1523/JNEUROSCI.0692-17.2017.

M. M. Nowicka, M. J. Wójcik, I. Kotlewska, M. Bola, and A. Nowicka. The impact of self-esteem on the preferential processing of self-related information: Electrophysiological correlates of explicit self vs. other evaluation. *PLoS ONE*, 13(7):1–16, 2018. ISSN 19326203. doi: 10.1371/journal.pone.0200604.

J. P. O'Doherty, P. Dayan, K. Friston, H. Critchley, and R. J. Dolan. Temporal difference models and reward-related learning in the human brain. *Neuron*, 38(2):329–337, 2003. ISSN 08966273. doi: 10.1016/S0896-6273(03)00169-7.

J. B. Overmier and M. E. Seligman. Effects of Inescapable Shock Upon Subsequent Escape and Avoidance Responding. *Journal of Comparative and Physiological Psychology*, 63(1):28–33, 1967. ISSN 00219940. doi: 10.1037/h0024166.

A. Padilla, C. Padilla, T. Ketterer, and D. Giacalone. Inescapable shocks and subsequent escape/avoidance conditioning in goldfish, Carassius auratus. *Psychonomic Science*, 20(5):295–296, 1970.

D. E. Parvin, S. D. McDougle, J. A. Taylor, and R. B. Ivry. Credit assignment in a motor decision making task is influenced by agency and not sensory prediction errors. *Journal of Neuroscience*, 38(19):4521–4530, 2018. ISSN 15292401. doi: 10.1523/JNEUROSCI.3601-17.2018.

V. Pawlak and J. N. Kerr. Dopamine receptor activation is required for corticostriatal spike-timing-dependent plasticity. *Journal of Neuroscience*, 28(10):2435–2446, 2008. ISSN 02706474. doi: 10.1523/JNEUROSCI.4402-07.2008.

V. Pawlak, J. R. Wickens, A. Kirkwood, and J. N. Kerr. Timing is not everything: Neuromodulation opens the STDP gate. *Frontiers in Synaptic*

Neuroscience, 2(OCT):1–14, 2010. ISSN 16633563. doi: 10.3389/fnsyn.2010.00146.

G. Peeters and J. Czapinski. Positive-negative asymmetry in evaluations: The distinction between affective and informational negativity effects. *European Review of Social Psychology*, 1(1):33–60, 1990. ISSN 1479277X. doi: 10.1080/14792779108401856.

C. Peterson. The meaning and measurement of explanatory style. *Psychological Inquiry*, 2(1):1–10, 1991.

C. Peterson and M. Seligman. Causal explanations as a risk factor for depression. *Psychological Review*, 91(3):347–374, 1984. URL `https://pdfs.semanticscholar.org/f9d1/0fb31c2726a4d92a12339f3e2b480be25ae2.pdf`.

C. Peterson, A. Semmel, C. Von Baeyer, L. Y. Abramson, G. I. Metalsky, and M. E. Seligman. The attributional style questionnaire. *Cognitive therapy and research*, 6(3):287–299, 1982.

P. K. Presson and V. A. Benassi. Locus of control orientation and depressive symptomatology: A meta-analysis. *Journal of Social Behavior and Personality*, 11(1):201–212, 1996. ISSN 08861641.

P. K. Presson and V. A. Benassi. Are depressive symptoms positively or negatively associated with the illusion of control? *Social Behavior and Personality*, 31(5):483–495, 2003. ISSN 03012212. doi: 10.2224/sbp.2003.31.5.483.

P. K. Presson, S. C. Clark, and V. A. Benassi. The levenson locus of control scales: Confirmatory factor analyses and evaluation. *Social Behavior and Personality: an international journal*, 25(1):93–103, 1997.

D. G. Rand and M. A. Nowak. Human cooperation. *Trends in Cognitive Sciences*, 17(8):413–425, 2013. ISSN 13646613. doi: 10.1016/j.

tics.2013.06.003. URL http://dx.doi.org/10.1016/j.tics.2013.06.003.

D. G. Rand, A. Peysakhovich, G. T. Kraft-Todd, G. E. Newman, O. Wurzbacher, M. A. Nowak, and J. D. Greene. Social heuristics shape intuitive cooperation. *Nature Communications*, 5:1–12, 2014. ISSN 20411723. doi: 10.1038/ncomms4677. URL http://dx.doi.org/10.1038/ncomms4677.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

R. A. Rescorla. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Current research and theory*, pages 64–99, 1972.

J. N. Reynolds and J. R. Wickens. Dopamine-dependent plasticity of corticostriatal synapses. *Neural Networks*, 15(4-6):507–521, 2002. ISSN 08936080. doi: 10.1016/S0893-6080(02)00045-X.

C. P. Richter. On the phenomenon of sudden death in animals and man. *Psychosomatic medicine*, 19(3):191–198, 1957. ISSN 00333174. doi: 10.1097/00006842-195705000-00004.

L. Rigoux, K. E. Stephan, K. J. Friston, and J. Daunizeau. Bayesian model selection for group studies - Revisited. *NeuroImage*, 84:971–985, 2014. ISSN 10959572. doi: 10.1016/j.neuroimage.2013.08.065. URL http://dx.doi.org/10.1016/j.neuroimage.2013.08.065.

C. J. Robins. Attributions and Depression: Why Is the Literature So Inconsistent? *Journal of Personality and Social Psychology*, 54(5):880–889, 1988. ISSN 00223514. doi: 10.1037/0022-3514.54.5.880.

C. J. Robins and A. M. Hayes. The role of causal attributions in the pre-

diction of depression. In G. M. Buchanan and M. E. Seligman, editors, *Explanatory Style*, pages 71–97. Routledge, 1995.

R. W. Robins, M. D. Spranca, and G. A. Mendelsohn. The Actor-Observer Effect Revisited: Effects of Individual Differences and Repeated Social Interactions on Actor and Observer Attributions. *Journal of Personality and Social Psychology*, 71(2):375–389, 1996. ISSN 00223514. doi: 10.1037/0022-3514.71.2.375.

D. M. Romney. Cross-validating a causal model relating attributional style, self-esteem, and depression: An heuristic study. *Psychological Reports*, 74 (1):203–207, 1994.

M. Rosenberg. Society and the adolescent self-image [dissertation]. *NJ: Princeton Univ*, 1965.

D. Roth and L. P. Rehm. Relationships among self-monitoring processes, memory, and depression. *Cognitive Therapy and Research*, 4(2):149–157, 1980. ISSN 01475916. doi: 10.1007/BF01173646.

R. H. Rozensky, L. P. Rehm, G. Pry, and D. Roth. Depression and self-reinforcement behavior in hospitalized patients. *Journal of Behavior Therapy and Experimental Psychiatry*, 8(1):35–38, 1977. ISSN 00057916. doi: 10.1016/0005-7916(77)90102-1.

M. F. Rushworth, M. A. P. Noonan, E. D. Boorman, M. E. Walton, and T. E. Behrens. Frontal Cortex and Reward-Guided Learning and Decision-Making. *Neuron*, 70(6):1054–1069, 2011. ISSN 08966273. doi: 10.1016/j.neuron.2011.05.014. URL `http://dx.doi.org/10.1016/j.neuron.2011.05.014`.

D. J. Russo, B. Van Roy, A. Kazerouni, I. Osband, and Z. Wen. A tutorial on Thompson sampling. *Foundations and Trends in Machine Learning*, 11(1): 1–96, 2018. ISSN 19358245. doi: 10.1561/2200000070.

H. Sadeghiyeh, S. Wang, M. R. Alberhasky, H. M. Kyllo, A. Shenhav, and R. C. Wilson. Temporal discounting correlates with directed exploration but not with random exploration. *Scientific Reports*, 10(1):1–10, 2020. ISSN 20452322. doi: 10.1038/s41598-020-60576-4. URL `http://dx.doi.org/10.1038/s41598-020-60576-4`.

D. P. Schmitt and J. Allik. Simultaneous administration of the Rosenberg self-esteem scale in 53 nations: Exploring the universal and culture-specific features of global self-esteem. *Journal of Personality and Social Psychology*, 89(4):623–642, 2005. ISSN 00223514. doi: 10.1037/0022-3514.89.4.623.

W. Schultz, P. Apicella, and T. Ljungberg. Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience*, 13(3):900–913, 1993. ISSN 02706474. doi: 10.1523/jneurosci.13-03-00900.1993.

W. Schultz, P. Dayan, and P. R. Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, 1997. ISSN 00368075. doi: 10.1126/science.275.5306.1593.

G. Schwartz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

M. E. Seligman. Learned helplessness. *Annual review of medicine*, 23(1): 407–412, 1972.

M. E. Seligman and S. F. Maier. Failure To Escape Traumatic Shock. *Journal of Experimental Psychology*, 74(1):1–9, 1967. ISSN 00221015. doi: 10.1037/h0024514.

M. E. Seligman, R. A. Rosellini, and M. J. Kozak. Learned helplessness in the rat: time course, immunization, and reversibility. *Journal of comparative and physiological psychology*, 88(2):542, 1975.

B. Seymour, N. Daw, P. Dayan, T. Singer, and R. Dolan. Differential encoding of losses and gains in the human striatum. *Journal of Neuroscience*, 27(18): 4826–4831, 2007. ISSN 02706474. doi: 10.1523/JNEUROSCI.0400-07.2007.

T. Sharot, C. W. Korn, and R. J. Dolan. How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience*, 14(11):1475–1479, 2011. ISSN 10976256. doi: 10.1038/nn.2949. URL `http://dx.doi.org/10.1038/nn.2949`.

S. Siegel and L. G. Allan. The widespread influence of the Rescorla-Wagner model. *Psychonomic Bulletin and Review*, 3(3):314–321, 1996. ISSN 10699384. doi: 10.3758/BF03210755.

B. Skinner. Superstition in the Pigeon (Reprinted From Journal Experimental Psychol, Vol 38, Pg 168-172, 1948), 1992. ISSN 0096-3445.

K. E. Stephan, W. D. Penny, J. Daunizeau, R. J. Moran, and K. J. Friston. Bayesian model selection for group studies. *NeuroImage*, 46(4):1004–1017, 2009. ISSN 10538119. doi: 10.1016/j.neuroimage.2009.03.025.

G. P. Strauss, M. J. Frank, J. A. Waltz, Z. Kasanova, E. S. Herbener, and J. M. Gold. Deficits in positive reinforcement learning and uncertainty-driven exploration are associated with distinct aspects of negative symptoms in schizophrenia. *Biological Psychiatry*, 69(5):424–431, 2011. ISSN 00063223. doi: 10.1016/j.biopsych.2010.10.015. URL `http://dx.doi.org/10.1016/j.biopsych.2010.10.015`.

L. P. Sugrue, G. S. Corrado, and W. T. Newsome. Matching behavior and the representation of value in the parietal cortex. *Science*, 304(5678):1782–1787, 2004. ISSN 00368075. doi: 10.1126/science.1094765.

R. S. Sutton and A. G. Barto. *Reinforcement learning: an introduction*. The MIT Press, 2018.

P. D. Sweeney, K. Anderson, and S. Bailey. Attributional Style in Depression. A Meta-Analytic Review. *Journal of Personality and Social Psychology*, 50(5):974–991, 1986. ISSN 00223514. doi: 10.1037/0022-3514.50.5.974.

R. W. Tafarodi and P. Walters. Individualism?collectivism, life events, and self?esteem: a test of two trade?offs. *European Journal of Social Psychology*, 29(56):797–814, 1999. ISSN 00462772. doi: 10.1002/(sici)1099-0992(199908/09)29:5/6⟨797::aid-ejsp961⟩3.3.co;2-j.

Y. Takikawa, R. Kawagoe, and O. Hikosaka. A possible role of midbrain dopamine neurons in short- and long-term adaptation of saccades to position-reward mapping. *Journal of Neurophysiology*, 92(4):2520–2529, 2004. ISSN 00223077. doi: 10.1152/jn.00238.2004.

S. E. Taylor. Asymmetrical effects of positive and negative events: the mobilization-minimization hypothesis. *Psychological bulletin*, 110(1):67, 1991.

H. Tennen and S. Herzberger. Depression, Self-Esteem, and the Absence of Self-Protective Attributional Biases. *Journal of Personality and Social Psychology*, 52(1):72–80, 1987. ISSN 00223514. doi: 10.1037/0022-3514.52.1.72.

H. Tennen, S. Herzberger, and H. F. Nelson. Depressive Attributional Style: The Role of Self-Esteem. *Journal of Personality*, 55(4):631–660, 1987. ISSN 14676494. doi: 10.1111/j.1467-6494.1987.tb00456.x.

E. Thomas and A. Baiter. Learned helplessness: Amelioration of symptoms by cholinergic blockade of the septum. *Science*, 1974.

S. C. Thompson. Will it hurt less if I can control it? A complex answer to a simple question. *Psychological Bulletin*, 90(1):89–101, 1981. ISSN 00332909. doi: 10.1037/0033-2909.90.1.89.

W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

W. R. Thompson. On the theory of apportionment. *American Journal of Mathematics*, 57(2):450–456, 1935.

E. L. Thorndike. A proof of the law of effect. *Science*, 77(1989):173–175, 1933.

W. S. Tillman and C. S. Carver. Actors' and observers' attributions for success and failure: A comparative test of predictions from Kelley's cube, self-serving bias, and positivity bias formulations. *Journal of Experimental Social Psychology*, 16(1):18–32, 1980. ISSN 10960465. doi: 10.1016/0022-1031(80)90033-5.

Z. Uludag. *Experimental Investigations on risk taking in sensorimotor decision making*. PhD thesis, School of Psychology, University of Leeds, 2019. URL `http://etheses.whiterose.ac.uk/27106/`.

C. Vázquez. Judgment of Contingency: Cognitive Biases in Depressed and Nondepressed Subjects. *Journal of Personality and Social Psychology*, 52 (2):419–431, 1987. ISSN 00223514. doi: 10.1037/0022-3514.52.2.419.

S. Venkatesh, M. L. Moulds, and C. J. Mitchell. Testing for Depressive Realism in a Clinically Depressed Sample. *Behaviour Change*, 35(2):108–122, 2018. ISSN 20497768. doi: 10.1017/bec.2018.12.

R. R. Vickers Jr, T. L. Conway, and M. A. Haight. Association between levenson's dimensions of locus of control and measures of coping and defense mechanisms. *Psychological Reports*, 52(1):323–333, 1983.

N. Vulkan. An Economist 's Perspective on Probability Matching. *Journal of Economic Surveys*, 14(7):101–118, 2000. ISSN 0950-0804.

F. H. Walkey. Internal Control, Powerful Others, and Chance: A Confirmation of Levensons Factor Structure. *Journal of Personality Assessment*, 43(5): 532–535, 1979. ISSN 15327752. doi: 10.1207/s15327752jpa4305_17.

M. E. Walton, T. E. Behrens, M. J. Buckley, P. H. Rudebeck, and M. F. Rushworth. Separable Learning Systems in the Macaque Brain and the Role of Orbitofrontal Cortex in Contingent Learning. *Neuron*, 65(6):927–939, 2010. ISSN 08966273. doi: 10.1016/j.neuron.2010.02.027. URL `http://dx.doi.org/10.1016/j.neuron.2010.02.027`.

S. Watanabe. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571–3594, 2010. ISSN 15324435.

S. Watanabe. A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, 14(1):867–897, 2013. ISSN 15324435.

WHO. *The global burden of disease: 2004 update*. World Health Organization, 2008.

R. C. Wilson and Y. Niv. Inferring relevance in a changing world. *Frontiers in Human Neuroscience*, 5(JANUARY 2012):1–14, 2012. ISSN 16625161. doi: 10.3389/fnhum.2011.00189.

P. T. Wong and B. Weiner. When people ask ”why” questions, and the heuristics of attributional search. *Journal of Personality and Social Psychology*, 40(4):650–663, 1981. ISSN 00223514. doi: 10.1037/0022-3514.40.4.650.

J. Wrase, T. Kahnt, F. Schlagenhauf, A. Beck, M. X. Cohen, B. Knutson, and A. Heinz. Different neural systems adjust motor behavior in response to reward and punishment. *NeuroImage*, 36(4):1253–1262, 2007. ISSN 10538119. doi: 10.1016/j.neuroimage.2007.04.001.

J. Yacubian, J. Gläscher, K. Schroeder, T. Sommer, D. F. Braus, and C. Büchel. Dissociable systems for gain- and loss-related value predic-

tions and errors of prediction in the human brain. *Journal of Neuroscience*, 26(37):9530–9537, 2006. ISSN 02706474. doi: 10.1523/JNEUROSCI.2915-06.2006.

S. Yagishita, A. Hayashi-Takagi, G. C. Ellis-Davies, H. Urakubo, S. Ishii, and H. Kasai. A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science*, 345(6204):1616–1620, 2014. ISSN 10959203. doi: 10.1126/science.1255514.

M. Zuckerman. Attribution of success and failure revisited, or: The motivational bias is alive and well in attribution theory. *Journal of Personality*, 47(2):245–287, 1979. ISSN 14676494. doi: 10.1111/j.1467-6494.1979.tb00202.x.