Author:
Stuchly, Erik

Title:
Examining the impact of reward landscape modulation on decision threshold selection

Examining the impact of reward landscape modulation on decision threshold selection

By

ERIK STUCHLÝ

School of Psychological Science
UNIVERSITY OF BRISTOL

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of MASTER OF SCIENCE BY RESEARCH in the Faculty of Life Sciences.

SEPTEMBER 2021

Word count: 18993

Abstract

The accumulation-to-threshold framework provides the best description of behaviour in perceptual decision-making tasks. Yet a long-standing question is how people select the appropriate decision thresholds. A popular hypothesis states that individuals treat sequential decision-making tasks as an optimisation problem, aiming to select policies which maximise reward rate. However, recent research shows that decision thresholds selected by participants are frequently sub-optimal. Instead of abandoning the optimisation hypothesis in response to this observation, it has been suggested that one cause of sub-optimal threshold selection is the distribution of reward rate across the threshold parameter space ('the reward landscape'); that is, rather than selecting the thresholds which yield optimal reward rate, individuals mainly avoid regions of the reward landscape which yield low rewards and select their thresholds from clusters of policies that yield high, albeit sub-optimal reward rate. The current project aimed to test this hypothesis, by first identifying whether it is possible to change the reward rate distribution within the reward landscape, while keeping the optimal policy the same. A systematic manipulation of task parameters of a simulated decision task revealed that manipulating the temporal or monetary penalty parameters modulated the reward rate distribution around the optimal policy. In a subsequent experiment, the monetary penalty levels in a decision task were manipulated, to test whether changing the reward landscape would affect the policies chosen by participants. In line with the expectations, decision-makers employed sub-optimal decision thresholds; however, there was no evidence of systematic threshold modulation as a function of reward rate distribution. Additionally, participants adopted thresholds further away from the optimal policy in a condition where this deviation caused the greatest loss of reward rate. These results suggest that decision-makers do not systematically explore the reward landscape when selecting a decision policy.

Dedication and acknowledgements

I would like to thank my supervisors, Dr Casimir Ludwig and Dr Gaurav Malhotra, whose contribution at every stage of this project was invaluable, and who helped make the past year a truly enriching and enjoyable experience for me.

Author's declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: .....ERIK STUCHLÝ..... DATE: .....28/09/2021.....

Table of Contents

List of Tables

TABLE **Page**

List of Figures

Chapter 1

General Introduction

Computational models of cognition attempt to describe behaviour through a set of underlying mechanisms, represented as adjustable parameters. This makes them a valuable tool in cognitive research, since fitting these models to data can explain which specific mechanisms are responsible for the observed behavioural patterns. However, an often unaddressed question is *why* these model parameters end up taking the values they do. A common, implicit assumption of many models is that agents are trying to optimize some outcome variable, where 'being optimal' simply means that the selected parameter values yield the best outcome out of all possible alternatives. For instance, the process of decision-making usually involves the selection from a range of options, some of which are more rewarding or otherwise superior to others. As such, when examining the principles that guide decision-making, any changes in model parameters are often interpreted in light of the assumption that participants adjusted their behaviour to optimise the outcome variable (such as reward).

When trying to understand the basic mechanisms that underlie decision behaviour, it might be tempting to study the so-called value-based decisions; for instance, economic decisions about monetary rewards (Tversky & Kahneman, 1974) or preferential decisions which involve incentives such as food (Padel & Foster, 2005). However, because these types of decisions are based on computing and comparing the subjective values of the offered options (Schultz, 2006), such decisions can be biased by a multitude of factors which affect the subjective valuation process. In contrast, decisions can also be made between 'value-neutral' alternatives, such as whether a cloud of dots on the screen moves with greater coherence to the left or to the right (de Bruyn & Orban, 1988), or determining whether the flashing circular grating is rotated clockwise or counter-clockwise (Kahnt et al., 2011). Relative to value-based decisions, these kinds of decisions are less likely to be affected by the decision-maker's specific utility functions, current needs or other subjective factors (White et al., 2012). In other words, perceptual decisions can be viewed as more objective because they reduce a major source of subjective variance, thus capturing de-

cision behaviour at a more fundamental level. That is why perceptual decision tasks are often the preferred paradigm in model-based studies of lower-level decision mechanisms (e.g. Summerfield et al., 2011).

1.1 Evidence accumulation models

Experiments on perceptual decision making typically involve a choice between two available options, one of which is objectively more correct or rewarding. A class of models known as evidence accumulation models (EAM) has been used extensively to account for measures of decision-making performance such as reaction times or accuracy, with a large degree of success (Smith & Ratcliff, 2004). Although multiple families of EAMs exist (linear ballistic accumulator, Brown & Heathcote, 2008; urgency gating, Cisek et al., 2009), they share the same core principle: that evidence from the environment is accumulated until it reaches a certain level - a decision threshold - which triggers a choice of the corresponding decision alternative.

Perhaps the most well-known type of EAM is the drift-diffusion model (DDM; Ratcliff, 1978), which captures decisions between two mutually exclusive alternatives. Like other models, the underlying principle of DDM is that decision makers accumulate noisy evidence based on samples obtained from the environment. This model further assumes two thresholds which lie on the opposite sides of the starting point of the evidence accumulation process, one for each decision alternative. Depending on which threshold the accumulated evidence reaches first, the corresponding option is chosen. What makes this model particularly useful for cognitive research is that each parameter represents a latent underlying psychological process; for example, the height of decision boundaries can be thought of as response caution, whereas non-decision time represents processes unrelated to decision itself, such as sensory encoding and motor execution (Schubert et al., 2016). Thus, fitting the model to data from different decisions can reveal which parameter values change and, by extension, which psychological processes are likely to be responsible for any differences in decision behaviour.

Although manipulating other DDM parameters (e.g. drift rate, non-decision time) allows the model to fit behavioural data from across different contexts, decision-makers have the most control over a single parameter - the height of decision thresholds. Studies show that individuals can increase or decrease the height of these boundaries according to task demands, so that more/less evidence respectively is required to trigger a decision (Franks et al., 2003). For instance, if one is instructed to prioritize response speed, the desired outcome can be achieved by lowering the height of the thresholds, enabling one to make a decision that is based on lower amounts of accumulated evidence (Herz et al., 2017). As Figure 1.1 shows, manipulating this single parameter enables the decision maker to switch between cautious and speedy response strategies - a well-established effect dubbed 'the speed-accuracy trade-off' (Heitz, 2014).

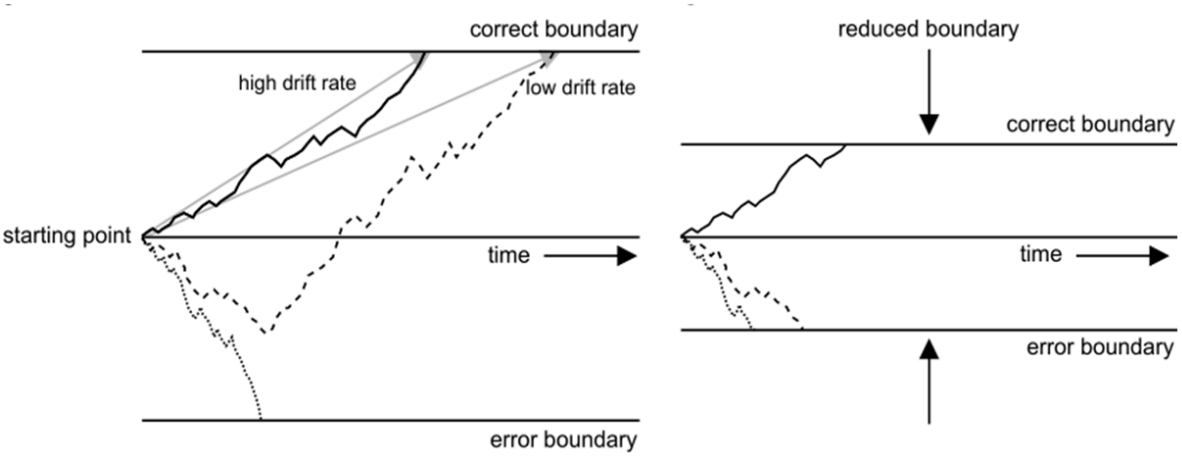

FIGURE 1.1. A visual representation of the speed-accuracy trade-off. If the decision boundaries are lowered, response times will be reduced but the risk of responding incorrectly increases, particularly for decisions with low drift rate (high uncertainty/difficulty). Taken from Lin et al. (2020)

In other words, the height of the decision thresholds determines when to wait for more evidence and when to make a decision, based on the levels of accumulated evidence and/or the elapsed time. In dynamic programming (Kaufman & Howard, 1961) and reinforcement learning (e.g. Sutton & Barto, 1998), a 'policy' $Q(x,t)$ is the probability distribution of all possible actions from the current state (x,t). In that sense, the decision threshold height can also be thought of as dictating the probability of taking the 'wait' or 'decide' action, given the current states of elapsed time and accumulated evidence. That is why the thresholds employed by a decision maker are often referred to as 'decision policy' (Drugowitsch et al., 2012), which is what thresholds will sometimes be called in the present text as well.

For most decisions, there is a specific set of thresholds that maximises accuracy while minimising response speed, allowing an individual to achieve the best possible performance (Bogacz et al., 2006) - the optimal policy with respect to some outcome variable. Multiple investigations have confirmed that decision makers are capable of trading off speed and accuracy of their responses optimally, in order to maximise gains on the task (e.g. Drugowitsch et al., 2015; Simen et al., 2009). Individuals' ability to select the optimal policy is enhanced when they are given training (Balci et al., 2011) and performance feedback (Evans & Brown, 2017), or when the inter-trial interval between two consecutive decisions is sufficiently large (Evans, Bennett, et al., 2019). These findings further demonstrate that decision-makers are able to maximize reward by inferring and selecting the optimal decision policy, and that they benefit from contexts conducive to efficient strategy learning.

1.2 Collapsing thresholds

Although the DDM successfully described decision-making processes across many contexts (Boehm et al., 2020), there is increasingly more evidence that one of its assumptions - decision thresholds being constant throughout the evidence accumulation period - is not always true. In order to maximise reward on certain tasks, the more beneficial approach is to gradually change the boundaries within a single decision. For example, decreasing boundaries represent the optimal policy when a decision deadline is imposed (Frazier & Yu, 2008): as the temporal deadline approaches, it is often more advantageous to make a potentially incorrect decision based on partial evidence, than to forfeit the chance to respond because time has run out (Miletić & van Maanen, 2019).

Similar examples can be found in situations where the aim is to maximise reward rate - reward per unit time. In these contexts, participants have a limited amount of time in which they can complete as many or as few decisions as they prefer. Due to the additional time constraints, balancing time and accuracy of responses is even more important if one wants to maximise the total reward obtainable within the available time-frame. For instance, on trials with highly informative evidence (low difficulty), it is optimal to adopt constant thresholds, since the probability of responding correctly is high even after observing low amounts of evidence. On the other hand, on trials with completely non-informative evidence (high difficulty), observing more evidence would not lead to a more informed decision. As such, the optimal strategy is to guess without observing any evidence - to collapse the thresholds immediately, or adopt constant thresholds with the height of zero. However, when easy and difficult trials are randomly intermixed and the decision-maker does not know which difficulty condition the next trial belongs to, the optimal strategy is an 'in between' approach - setting the thresholds rather high initially to avoid making early errors, and gradually collapsing them to minimize time spent on the decision if that particular trial turns out to be one with uninformative evidence (Malhotra et al., 2018; Palestro et al., 2018; Moran, 2014). In other words, adopting collapsing thresholds is of particular benefit when there is uncertainty about vital aspects of the decision process, such as the quality of evidence or the amount of remaining time.

Nevertheless, there is an ongoing debate about whether individuals actually employ collapsing thresholds, with certain authors concluding that such policies are an unnecessarily complicated construct which rarely improves a model's fit to decision-making data (Voskuilen et al., 2016). The most comprehensive study of the phenomenon conducted so far is a meta-analysis which looked at 9 datasets from 8 separate studies of perceptual decision-making (Hawkins et al., 2015). When pooled across studies, decision behaviour of half the monkey subjects was better explained by the the collapsing thresholds model than the constant thresholds model variant. However, data from human decision-makers was overwhelmingly better described by

the constant thresholds models, suggesting that people do not systematically employ collapsing thresholds. It is, then, not too surprising that the very concept of time-varying decision thresholds remains controversial in the field.

However, most of the studies included in Hawkins et al. and beyond (e.g. Voskuilen et al., 2016) failed to check whether collapsing boundaries represent the optimal decision policy on a given task. As a representative example, Voskuilen et al. fitted DDM variants with constant and collapsing boundaries to behavioural data from several classical decision tasks, such as the perceptual dot motion discrimination task (de Bruyn & Orban, 1988). They discovered that the model specification with constant boundaries provided a somewhat better fit to data than the one which assumed time-varying thresholds. This is hardly surprising though, given that the only experimental manipulation was whether the instructions encouraged speed or accuracy of the response - a manipulation which has been shown to affect the height of decision boundaries adopted by participants, but not their degree of change within a single trial (Katsimpokis et al., 2020). Since the decision context did not promote the use of collapsing thresholds as the optimal policy, it stands to reason that a model with fixed boundaries would describe the data well; in fact, the authors themselves note that, in the collapsing thresholds version of the model, the degree of collapse was so small as to be qualitatively equivalent to constant thresholds. This intuition was confirmed by Malhotra et al. (2018), who computed the optimal boundaries for a number of aforementioned studies and found that constant or slightly decreasing boundaries represented the optimal strategy in many of the tasks.

On the other hand, studies which ensured that the decision context is conducive to selecting time-varying decision policies usually find that individuals, indeed, employ them. For example, Khodadadi et al. (2017) used an expanded judgment task, where an animated boat continuously moved left or right along the horizontal axis and participants had to determine which side the boat would eventually reach. Crucially, the decision trials were randomly intermixed such that the boat either moved in one direction most of the time (easy trials), or its direction of motion was near-random (difficult trials). In line with the optimality analysis predictions outlined above, most participants used collapsing decision thresholds throughout the experiment. Other studies confirmed these findings in context with intermixed decision difficulties (Malhotra et al., 2017), as well as other contexts where collapsing thresholds represent the optimal policy - such as when sampling costs within a trial increase with each new sample (Busemeyer & Rapoport, 1988), or when the quality of evidence changes within a trial (Derosiere et al., 2021). As such, it appears that individuals are capable of using time-varying thresholds when it is adaptive to do so.

1.3 Reward maximization via time-varying thresholds

The finding that individuals use time-varying policies when it is the optimal policy with respect to reward rate (reward per unit time) implies that decision-makers do this to maximise gains on the task. To an extent, this assumption is justified - as described earlier, individuals seem motivated to maximise their reward rate by searching for and selecting the optimal constant thresholds in standard decision contexts (Drugowitsch et al., 2015). But there is also evidence which contradicts this assumption. In a recently conducted decision experiment where collapsing thresholds represented the optimal policy, Evans, Hawkins, and Brown (2019) participants failed to adopt collapsing boundaries even with a fixed experimental time and the explicit instruction to optimise reward rate - conditions that should motivate the participants to optimize behaviour by selecting collapsing thresholds. Thus, the claim that individuals aim to optimize reward rate by adopting time-varying decision boundaries has not received unanimous support.

One potential drawback of studies such as Evans et al. is that they only used hypothetical reward, instead of real-life incentives such as money. Because people are generally more sensitive to real than imagined monetary rewards (Xu et al., 2018), it can be argued that participants had little incentive to maximize their reward rate. Indeed, in a typical multi-alternative decision task with imaginary rewards, Hawkins et al. (2012) showed that participants do not attempt to maximize reward rate, aiming instead for a certain acceptable level of accuracy while minimizing the time spent in the experiment. As such, the interpretation that individuals do not select the optimal time-varying policies due to lack of incentive cannot be ruled out.

In contrast, Malhotra et al. (2017) took stricter measures to ensure that participants aim to optimize reward rate. Similar to Evans, Hawkins, and Brown (2019), participants in this study completed a series of decisions within a limited total time and received the instruction to maximize gains. However, they were also told that the score accumulated across decisions would later get converted to money and be paid out to them, and that the best-performing participant would get paid an extra large bonus. The task itself consisted of observing a series of discrete Gabor patch flashes that could appear either on the left or right side of the screen, deciding which side the flashes occur on more often. The probability of the flash occuring on the 'correct' side for a given trial - the quality of evidence - varied across the types of decision blocks: high probability (easy-only block), near-chance probability (difficult-only block), or both types of trials intermixed randomly (mixed block).

As expected from the findings of Malhotra and colleagues' optimality analysis, participants largely employed thresholds with negative slopes on mixed experimental blocks; however, Figure 1.2 shows that participants used thresholds with a more negative gradient than what would be considered optimal (-30°as opposed to -15°). Moreover, participants also used collapsing boundaries in single-difficulty easy blocks, where the optimal policy was to use fixed boundaries. In

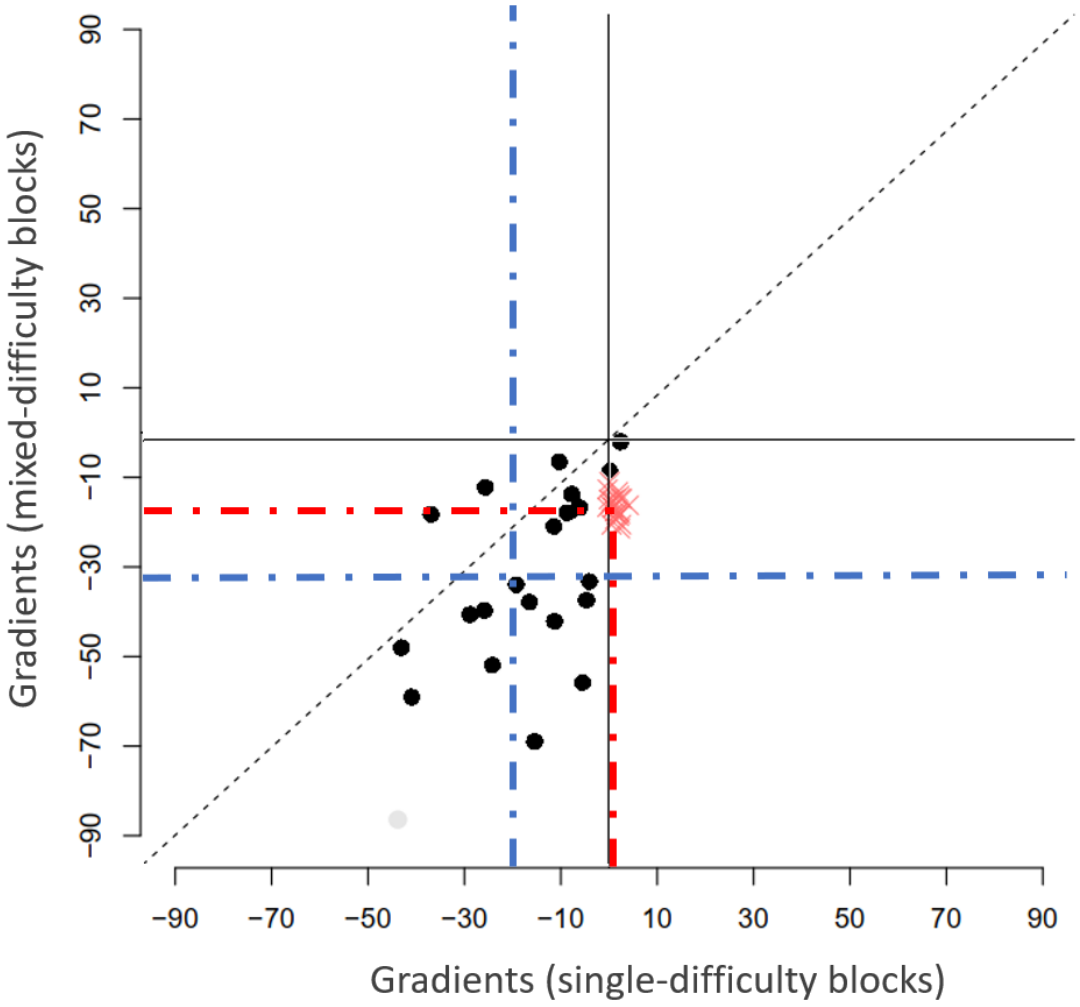

FIGURE 1.2. Participant-level threshold gradients used in single-difficulty easy and mixed-difficulty easy blocks in Malhotra et al. (2017). Red crosses represent the simulated, optimal agents, with black dots depicting individual-level gradient estimates. Means of the simulated participants' gradients are shown as the red lines, whereas blue lines represent mean gradients for the real participants. These indicate a tendency of real participants to 'over-collapse' decision thresholds in both types of difficulty blocks.

other words, participants over-collapsed their decision boundaries across experimental blocks - a pattern that is perplexing for two reasons. First, it contradicts the finding that, if decision makers use thresholds with sub-optimal gradient, it is typically in the direction of smaller/less negative gradients than optimal (Evans, Hawkins, & Brown, 2019; Hawkins et al., 2015). Second, decision-makers generally show an 'over-cautious bias' when selecting the height of the thresholds, manifested by setting decision boundaries too high (Balci et al., 2011). The act of over-collapsing one's boundaries seems counter-intuitive in light of this bias, since the decreasing height would promote quicker and potentially more error-prone decisions as the trial progresses.

To get a better understanding of this effect, Malhotra et al. went beyond the traditional approach of simply comparing the values of adopted against optimal threshold parameters - they also examined the levels of reward rate the adopted policies could yield. To this end, they generated a 'reward landscape' for each type of decision block - a heatmap showing the reward rate associated with each combination of initial threshold heights and gradients (Figure 1.3). Of course, the decision-maker does not know the reward rate distribution across different thresholds, which means they have to infer this information somehow. The idea is that the selection of appropriate threshold gradient resembles the selection of threshold heights on decision tasks, whereby participants try out different thresholds until they identify the best-performing policy (Simen et al., 2009). So, participants presumably select an initial set of decision thresholds for the task and then sample a range of decision policies (different height and gradient combinations), generally favouring policies that increase their reward rate and abandoning those which decrease it, until they obtain levels of reward they find acceptable. In that sense, participants are thought to act similar to 'hill climbing' optimisation algorithms (e.g. Kimura et al., 1995; Katayama et al., 2000), which make incremental, iterative changes to the current solution in order to find the optimal solution to a problem.

Due to the presumed need for decision-makers to try different policies, the amount of reward rate those policies yield could have significant impact on which one is ultimately chosen. For instance, the reward landscape from 'easy blocks' in Malhotra et al. displayed in Figure 1.3 is asymmetric around the cluster of optimal policies - the potential rewards decrease only gradually as one employs increasingly more negative gradients, while the drop in reward is considerably steeper as gradients become more positive. Figure 1.4 illustrates that, for a decision-maker who adjusts their response criterion if it leads to unsatisfactory outcomes (Myung & Busemeyer, 1989), it should be rather easy to reach the wide range of policies with negative gradient which yield near-optimal levels of reward rate (light red circle). Exploration within this area yields little increase in reward rate; however, when the decision-maker samples policies with a more positive gradient (pink circle), they will observe a rapid decrease in reward rate. This is likely to be perceived as a negative outcome (Wischnewski & Schutter, 2018), prompting the decision

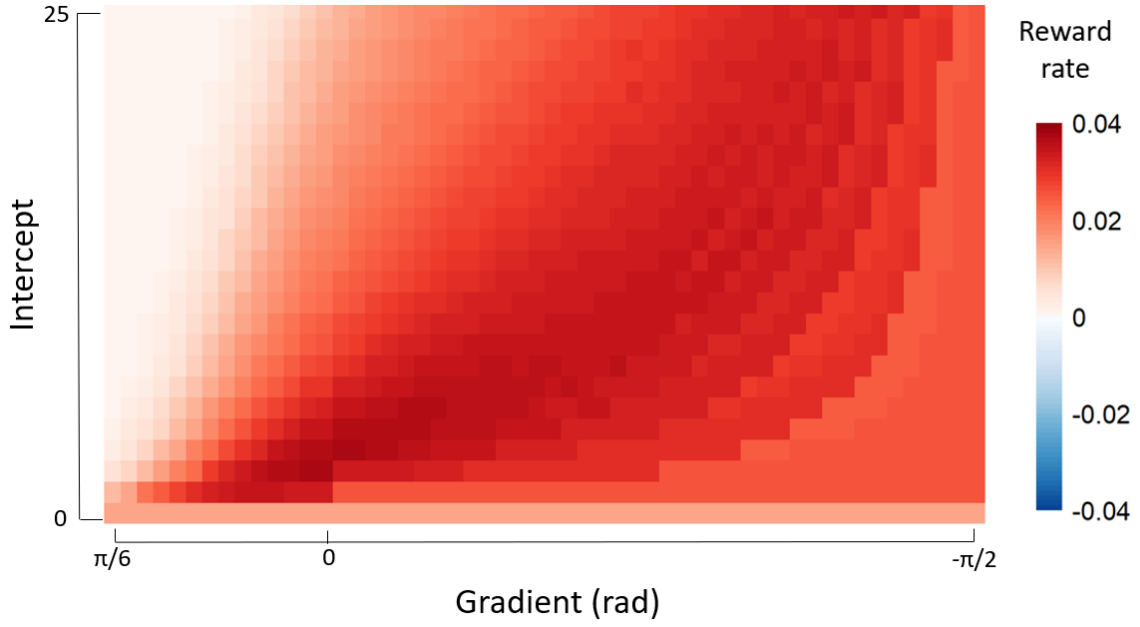

FIGURE 1.3. A heatmap depicting the reward landscape of single-difficulty easy decision blocks in Malhotra et al. (2017). Gradient values range from slightly increasing (positive) to instantly collapsing (negative). A cluster of optimal policies occurs around threshold parameter values of $gradient = 0$ and $intercept = 3$

maker to return to and possibly settle on one of the tested policies with a slightly negative gradient. Being discouraged from further exploration would mean that the decision-maker misses the opportunity to identify and reach the relatively narrow range of policies which yield the optimal reward rate (dark red circle). Thus, Malhotra and colleagues proposed that settling for these negative-gradient policies which yield a 'good enough' reward rate might be a risk-avoiding strategy, brought about by the dangers of selecting thresholds with a slightly increasing gradient. In this way, the 'shape' of the reward landscape - specifically, the asymmetry in reward rate around the optimal policy - could provide an intuitive explanation for the pattern of results observed in the Malhotra et al. experiments.

This proposed 'good enough' approach relates to the concept of satisficing, coined over half a century ago (Simon, 1955). The idea is that if an organism is unable to discern the optimal solution to a problem, then it will choose a solution which satisfies as many of its needs as possible. This approach is particularly useful if there is uncertainty about the problem or the optimal solution to that problem (Simon, 1972) - which is true for selection of decision policies, since individuals do not have prior knowledge of how much/little reward rate the different policies yield on the particular task. Moreover, the idea that individuals would 'satisfice' in order to

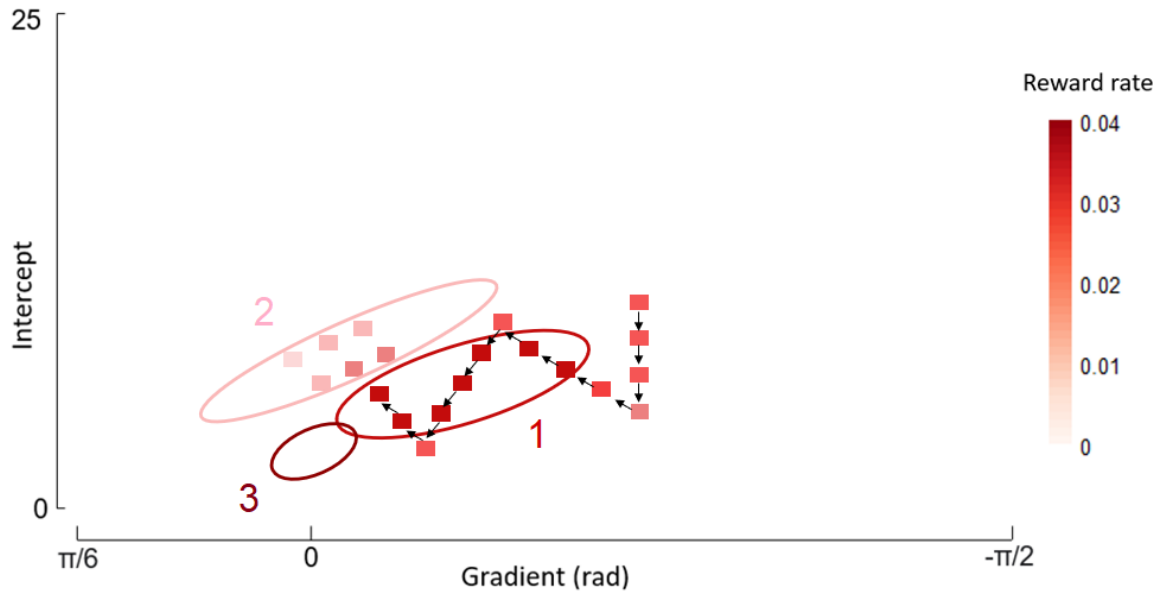

FIGURE 1.4. An illustration of reward landscape exploration from the participant's perspective. After selecting an initial policy and consequently adopting policies which increase reward rate, most participants will arrive at the policies with negative gradient which yield near-optimal reward rate (marked by elliptical region 1). However, the policies where reward drops rapidly are just adjacent to this relatively large area (region 2), meaning that further exploration poses a risk of losing reward. That is why participants may prematurely abandon the search for the most optimal policies (region 3) and settle for a policy which yields 'good enough' reward rate (region 1)

avoid risk is congruent with decision-making literature on risk aversion; individuals seem particularly motivated to actively avoid risks if a change to status quo could lead to worse outcomes (Huber et al., 2014). As such, this 'good enough' approach to selecting decision thresholds seems to be a viable explanation that deserves further scrutiny.

1.4 Current study

This explanation for the findings of Malhotra et al. (2017) was made post-hoc as a speculative account of the data, and so the evidence for it is merely correlational. However, the account allows for making potentially testable predictions. The central idea is that individuals used 'over-collapsing' thresholds on the task because of the asymmetry in reward rate distribution around the optimal policy, which favoured thresholds with negative gradients. As Figure 1.5 shows, reducing this asymmetry - for example, by decreasing the reward rate from negative-gradient policies (blue curve) - would result in thresholds with negative gradient no longer yielding the

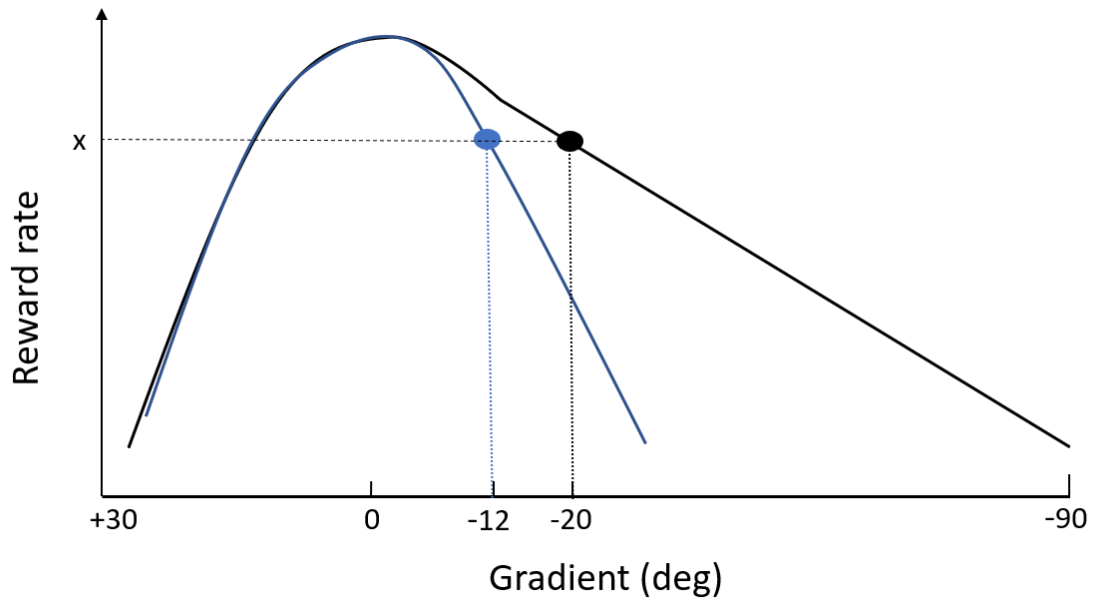

FIGURE 1.5. An imaginary example of how changing the reward landscape symmetry might affect the selection of decision policy. Black curve represents the cross section through the original reward landscape, whereas the blue curve shows a modified reward landscape which is more symmetrical around the point of optimal policy (gradient of zero).

same levels of reward rate (x). If the decision-maker wanted to retain the same levels of reward rate, they would have to adjust the gradient of their thresholds such that it is closer to the optimal policy (-12°as opposed to -20°). In other words, over-collapsing one's boundaries would no longer be a viable risk-avoiding strategy. Indeed, if the degree of asymmetry were reversed in favour of policies with a positive gradient, over-collapsing the boundaries would prove detrimental. As such, if individuals explore different policies and are sensitive to the reward rate these policies yield, we would predict that increasing the landscape symmetry should motivate participants to settle on policies closer to the optimal one. It is, however, presently unknown which aspects of the task need to be adjusted in order to change the reward landscape symmetry.

To address these problems and questions systematically, the current project aims to:

1. Build a simulation of the decision task and manipulate its parameters, to determine which ones (if any) produce a reward landscape with a different degree of symmetry around the optimal policy. (Computational part)

2. Manipulate the corresponding parameters on a decision-making task and compare the gra-

dients of thresholds used by participant across conditions, to investigate whether higher landscape symmetry promoted the choice of more optimal decision policy. (Experimental part)

Chapter 2

Changing the reward landscape symmetry

2.1 Introduction

Evidence accumulation models assume that a decision-maker samples and integrates evidence from the environment until a certain threshold is reached, at which point the corresponding decision is made. Recent studies found that in certain contexts, the optimal behaviour is achieved by using a time-varying, instead of constant thresholds. In these situations, the optimal policy typically requires that the height of these thresholds decreases over time, which is why they are often called collapsing boundaries (e.g. Ditterich, 2006). Whether decision-makers aim to optimise reward rate by using collapsing thresholds, however, remains uncertain, with certain studies finding evidence that individuals use constant thresholds even when collapsing boundaries represent the optimal policy (Evans, Hawkins, & Brown, 2019).

In contrast to these studies, Malhotra et al. (2017) found that participants in their experiments over-collapsed their decision boundaries across experimental conditions - instead of using threshold gradients which would yield the optimal reward rate, they consistently employed ones with relatively more negative gradients, obtaining sub-optimal levels of reward rate. Not only does this finding confirm that decision-makers can use collapsing thresholds when they represent the optimal policy (e.g. Khodadadi et al., 2017), but also shows that participants can over-use these policies to the point of behaving sub-optimally. In addition, this finding also contrasts the commonly observed bias of decision-makers for accuracy over speed in most classical tasks (Balci et al., 2011), making it even more surprising.

Unlike the previously mentioned studies, Malhotra and colleagues did not only compare the adopted thresholds to a single optimal policy for the task, but instead computed the reward rate associated with various combinations of decision threshold's gradient and intercept,

generating a three-dimensional 'reward landscape' upon which each individual's policy could be superimposed. The specific 'shape' of the reward landscape, whose cross-section is shown in Figure 2.1, offered a possible explanation for why participants over-used collapsing thresholds: relative to the optimal policy on the task, employing thresholds with a more positive gradient would result in a much greater decrease in reward rate than employing thresholds with more negative gradients. Under this view, participants were being risk-averse - to avoid accidentally using thresholds with a positive gradient and incurring large losses, they instead settled on thresholds with slightly negative gradients which yielded acceptable levels of reward rate. Provided this interpretation is true, then increasing the degree of this landscape asymmetry should result in participants employing a decision policy with closer-to-optimal gradient, since over-collapsing the thresholds would suddenly incur larger losses as well, making it less useful as a risk-avoiding policy. Before testing this hypothesis though, it is necessary to identify a way to change the reward landscape symmetry, as Malhotra et al. were seemingly the first to examine reward rate landscapes in decision-making contexts.

The aim of the simulation work presented here was to develop a computational simulation of the expanded judgment task from Malhotra et al. and systematically manipulate its parameters, to examine which, if any, have the capacity to generate a reward landscape with a different degree of symmetry around the optimal policy. Crucially, the desired landscape needed to satisfy two conditions:

1. The optimal policy remains largely the same across manipulations, and
2. The manipulation produces a change in reward rate distribution either for positive, or negative gradients exclusively.

The reason for requirement 1 is that the optimal policy would serve as an anchor point across manipulations, ensuring that participants are always trying to achieve the same goal when selecting their decision policy. Requirement 2 is somewhat less essential in comparison, but it was desirable to keep one side of the landscape (relative to the optimal policy) similar across manipulations, so that the risks or benefits associated with exploring those landscape regions would also remain similar. These two requirements would ensure that any observed change in policies adopted across manipulations could be attributed solely to changes in reward rate on one side of the landscape.

The following section outlines the process by which the simulation was constructed, as well as the findings obtained from systematically manipulating the simulated task parameters. Ultimately, parameters related to the penalty for incorrect decisions satisfied the two requirements of an ideal manipulation as outlined above, enabling this project to test its main hypothesis in Chapter 3.

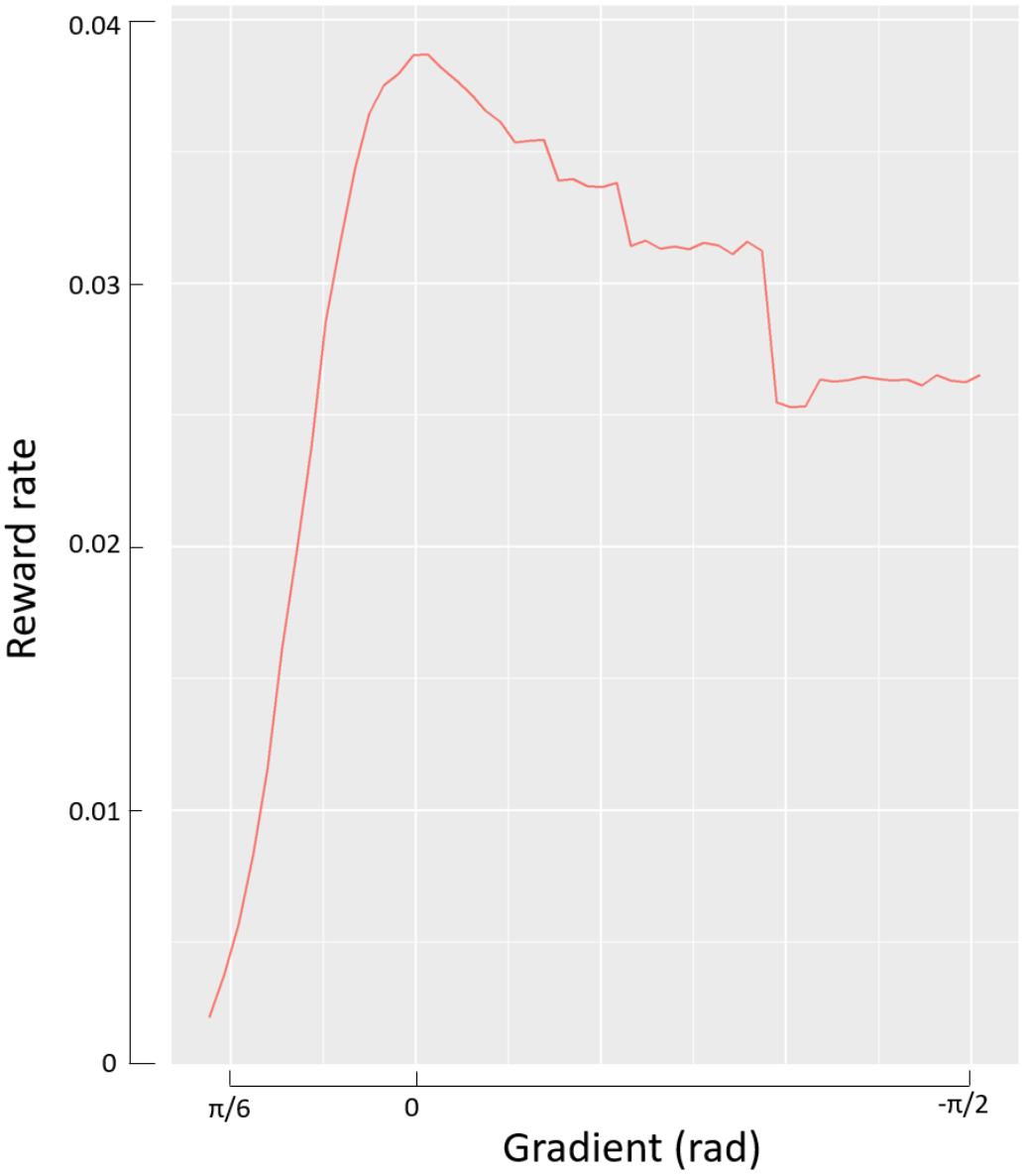

FIGURE 2.1. Reward rates for all gradient values across $intercept = 3$, from the single-difficulty easy blocks in Malhotra et al. (2017). The decrease in reward rate is much more rapid for positive than negative gradients, relative to the optimal policy of $gradient = 0$

2.2 The simulation exercise

2.2.1 The modelled task and its parameters

In order to generate a reward landscape and identify suitable parameter manipulations, it was necessary to model the expanded judgment task used in Experiments 2 and 3 of Malhotra et al. (2017). Broadly speaking, the task consisted of making a series of perceptual decisions within a fixed experimental time, meaning that the optimal decision policy was one that maximised accuracy while minimising response times - maximised reward rate, calculated as the sum of rewards divided by the total elapsed time. Within the time-limited decision block, participants completed a series of decision trials, which consisted of a sequence of briefly flashing decision samples (Gabor patches). Each sample in the sequence could appear either on the left or the right side of the screen and participants had to decide which side the samples appear on more often, by pressing a keyboard button at any point throughout the sample presentation period. If the probability of the sample appearing on either side was 0.5 (completely random), then the task was very difficult, whereas if these probabilities approached 1/0 for either side (appearing almost exclusively on one side), the task became significantly easier. As such, the probability of the stimulus appearing on either side of the screen was directly related to task difficulty; by extension, the difference between this probability and the random probability of 0.5 could be thought of as the **drift rate**, where high value indicates an easy task and low value indicates a difficult task.

Within the decision trial, each individual sample was presented for a fixed amount of time, also called the inter-stimulus interval (**ISI**). There was also a fixed period (inter-trial interval; **ITI**) between two consecutive decision trials, to ensure that participants have enough time to prepare for the upcoming trial. If the response on a particular trial was correct, participants gained a fixed amount of score (**reward**) and proceeded to the next decision after the ITI period. If the response was incorrect, no score was gained or lost (the **penalty** was zero); however, additional waiting time was added to the ITI period as an alternative form of penalty (ITI replaced with **ITIp**). Table 2.1 includes a full list of task parameters and their values as used in Malhotra et al., Experiment 2a, single-difficulty easy decision blocks, which provided basis for the current simulation. Thus, a task with this set of parameter values will be referred to as 'the default' condition in the remainder of this chapter.

2.2.2 Simulating 'the default' decision task

The simulation was implemented in base R with RStudio (R Core Team, 2019). The simulated decision process itself was based on a simple random-walk model (e.g. Litterman, 1983) and assumed that evidence accumulation occurs in a discrete two-dimensional time-evidence space. The general principle for each simulated trial is shown in Figure 2.2: both time (t) and evi-

Table 2.1: A list of task parameters and their values used in Malhotra et al. (2017), experiment 2a, 'Easy' trials

Parameter name	Value
Drift rate	0.2
ISI	200ms
ITI	3s
ITIp	7s
Reward	2p
Penalty	0p

dence (x) states were initially at zero. The arrival of a new evidence sample always changed the current time state t to t+1; the evidence state x was changed to either x+1 or x-1, depending on which decision alternative the evidence sample supported (the probabilities being equal to $0.5 \pm driftrate$). When the evidence state crossed the decision threshold at any particular time state, the response was coded as correct or incorrect (depending on which threshold was crossed) and both time and evidence states were re-set back to zero. This marked the end of the simulated trial - if the response was correct, a reward of 1 unit was added to the total score, and the ISI for each observed sample, as well as the ITI were added to the total elapsed time . On the other hand, the score remained unchanged if the decision alternative was incorrect - however, a time penalty equal to the value of ITIp parameter was added to the total elapsed time.

2.2.2.1 The decision thresholds

As mentioned before, the simulated decision process terminated with a response once the decision threshold was reached by the accumulated evidence. Because decisions were always made between two mutually exclusive alternatives, the simulation included a pair of evidence thresholds equidistant from the starting point, each corresponding to one decision alternative. It was assumed throughout that the threshold with positive intercept always triggered the correct option, while the negative one represented the incorrect option. The threshold values at each time point were determined by two parameters: intercept and gradient. The intercept was simply conceptualised as the starting value of the threshold – its height at state $t = 0$. The gradient was a value expressed in radians, whose tangent determined the amount of decrease in thresholds' height on each subsequent time step. In line with Malhotra and colleagues' dynamic programming method, simulations were carried out for all decision policies with intercepts between 0 and 25, and gradients between $\pi/6$ (slightly increasing) to $-\pi/2$ (instantly collapsing) in $-\pi/80rad$ steps - a range which includes all types of policies that could conceivably be used by decision makers on this kind of task.

It should be noted that the linear shape of thresholds is an assumption made mostly for convenience. Depending on the decision context, some researchers found that the most opti-

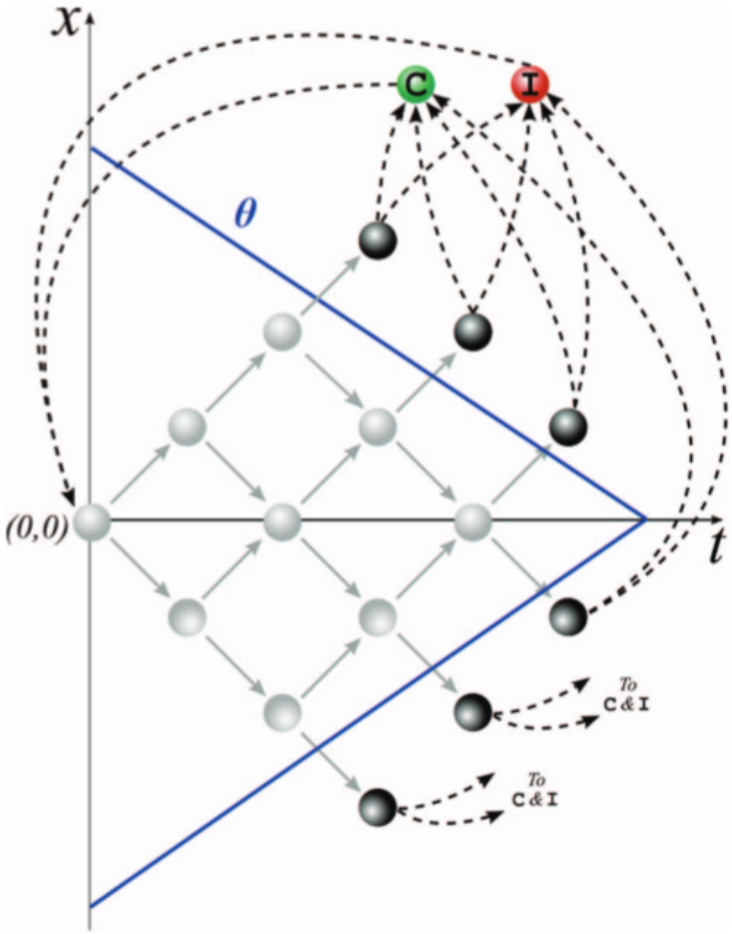

FIGURE 2.2. The decision process as a time-evidence space. Each grey ball represents a possible discrete time-evidence state, with x-axis representing time and y-axis the amount of accumulated evidence. Once a threshold is reached (blue line), a response is coded as either correct (C) or incorrect (I), the process returns to the initial state of (0,0) and a new trial commences. Illustration taken from Malhotra et al. (2017)

mal policy is represented by nonlinear collapsing boundaries, typically resembling a Weibull, S-shaped curve (Ratcliff et al., 2003). While potentially providing a better fit to data, the definition of these thresholds requires more than two parameters, making the comparison of reward landscapes for different policies significantly more challenging than assuming linear thresholds with only two parameters. Ultimately though, the specific shape of boundaries employed in the current simulation is of little consequence, because the optimal policy - constant boundaries - would look identical regardless of the framework used. As such, any deviation from this optimal policy as captured by the linear thresholds framework would also represent a sub-optimal policy in the Weibull thresholds framework.

On certain simulated trials, particularly with the combination of large intercepts and positive gradients, the evidence process never reached a threshold. To deal with this situation, it was necessary to introduce an upper limit at which the simulated trial terminated. The value of this limit was set to $n = 50$ samples - a high enough value to avoid the premature termination of a trial on decision policies which depend on a larger number of samples. When this limit was reached, the decision was encoded as incorrect, incurring time and, where applicable, monetary penalty, in addition to the time spent in the trial (i.e 50 samples). In that sense, this time-out rate could be thought of as an additional task parameter; however, since manipulation of this parameter had a relatively small effect unless its value changed drastically, it was simply fixed at 50 time units and all subsequent reward landscapes were calculated with this fixed time-out rate.

2.2.2.2 Calculating the reward rate

Given the set of task parameters listed in Table 2.1, the reward rate for a particular policy was calculated as follows:

$$RR = \frac{\sum_{i=1}^{n} reward_i - \sum_{i=1}^{n} penalty_i}{\sum_{i=1}^{n} (ITI_i + ITIp_i + ISI * latency_i)} \tag{2.1}$$

where i denotes the index of an individual simulated trial, and n represents the total number of simulated trials (set to 50000). In other words, reward rate was computed by dividing the total sum of rewards by the total time, across many simulated trials. The only value in these calculations which was not a set parameter - latency - was the number of samples observed before a threshold was reached and its value depended both on the current decision policy, as well as the specific sequence of evidence samples generated for trial i. To generate the reward landscape, this formula was used to calculate reward rate for each decision policy (intercept * gradient combination of the decision threshold).

Because reward rate given by equation 2.1 was expressed as reward per unit time, it was important to define both the values of reward (simply set to 1) and, more crucially, the time-related parameters. The assumption adopted in the simulation was that evidence is presented

at a rate of 1 sample per unit time. By definition, this means that the value of the ISI parameter was 1 time unit, and since the ISI in the Malhotra et al. experiment was 200ms, one time unit in the simulation corresponded to 200ms in real time units. All the other time-related parameters (ITI, ITIp) were then expressed as multiples of this unit time - for instance, because ITI was 15 times higher (3s) than ISI in the Malhotra et al. experiment, the value of ITI simulation parameter was set to 15 time units. Table 2.2 shows the values of all six task parameters for the default condition, expressed in the standardised units.

Table 2.2: A list of task parameters and their values used in the default version of the simulation

Parameter name	Value(standardised simulation units)	Value(real units)
Drift rate	0.2	0.5 ± 0.2
ISI	1	200ms
ITI	15	3000ms
ITIp	35	7000ms
Reward	1	2p
Penalty	0	0p

One consequence of the ISI parameter being assigned the value of 1 was that response times were coarsely quantised, encoded in terms of the number of evidence samples, rather than milliseconds. However, this relative lack of precision should not affect the reward rate computations in any significant way, with the only noticeable effect being somewhat sharper differences between reward rate values of adjacent policies in the reward landscape.

Another consequence of ISI being set to 1 (in addition to penalty being 0 in the default simulation) is that Equation 2.1 for calculating reward rate was effectively reduced to:

$$RR = \frac{\sum_{i=1}^{n} reward_i}{\sum_{i=1}^{n} (ITI_i + ITIp_i + latency_i)} \tag{2.2}$$

Because all parameters were defined in terms of these standardised units (reward units and time units), the reward rate was also expressed in the standardised form of 'units of reward per unit time'. Consequently, the actual reward rate for any given experiment would be a scalar multiple of this normalised reward rate, obtained by multiplying the values of reward and time-related parameters by the appropriate multipliers. For instance, if one were interested in calculating the average reward per second and assumed the reward to be 2p and one time unit to correspond to 200ms (as in the default task), then this actual reward rate would be calculated as:

$$RR_{actual} = RR_{normalised} * \frac{2p}{200ms * \frac{1}{1000}} = RR_{normalised} * 10p/second \tag{2.3}$$

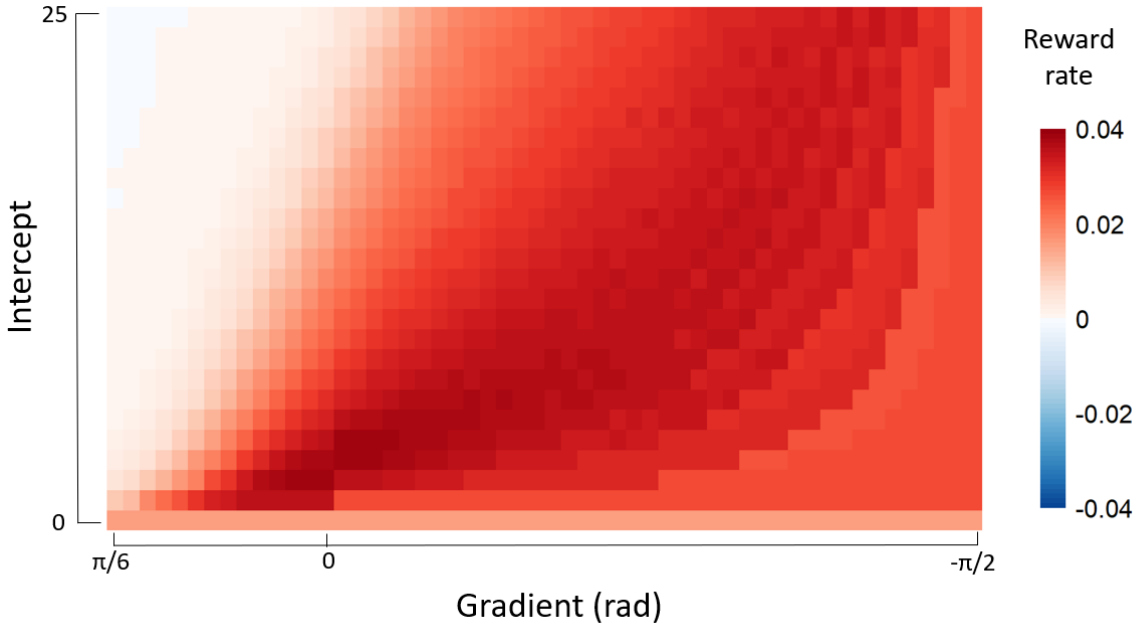

FIGURE 2.3. A heatmap depicting the reward landscape for the default condition, as generated by the simulated random-walk decision process. Just like in the landscape generated by Malhotra et al. (2017), the cluster of optimal policies occurs around threshold parameter values of $gradient = 0$ and $intercept = 3$, with a different degree of reward rate decrease to either side of this optimal policy.

2.2.2.3 The default reward landscape

Running the simulation as outlined above generated the normalised reward landscape displayed in Figure 2.3. On a qualitative level, it looks similar to the landscape computed by Malhotra et al via dynamic programming. For example, the optimal policy is the one with intercept = 3 and gradient = 0 ($RR_{normalised} \approx 0.04$), and there is a gradual decrease in reward rate at negative gradients, contrasting with the rapid decrease at positive gradients. A policy-level comparison of a reward landscape generated by the current simulation and one generated by the dynamic programming method employed in Malhotra et al. confirmed that both methods give highly similar landscapes, with the same optimal policies and near-identical distribution of reward rate across all policies (see Appendix A for more details). However, the simulation method is conceptually closer to the evidence accumulation framework which the current investigation is based upon, because it makes reward rate calculations based on the decision process directly. As such, all subsequently reported reward landscapes were generated by the simulation as described above.

To match the dynamic programming method used in computations of reward landscape in

Malhotra et al., the current simulations assumed an infinite time horizon and instead used a fixed number of decision trials (n=50000). However, the Malhotra et al. study and the experiment reported in the Experimental Section include a fixed total time, meaning that the total number of trials varied across sessions. Because the simulation calculates reward rate – the reward per unit time - rather than reward on its own, the total amount of experimental time should make little difference, since any number of trials large enough to minimize output noise should generate an accurate reward landscape. To examine whether this is the case, a slightly modified version of the simulation was run, which fixed the experimental time at 5 minutes instead (1500 time units; same time limit as employed in the Malhotra et al. experiments), terminating the decision process when this time limit was reached. In line with the prediction, running this simulation 10000 times and averaging the reward rates for each decision policy yielded a landscape identical to the simulation with an infinite time horizon. As such, all results presented from here on will be based on the simulation with $n = 50000$ trials.

To summarise, the following pseudo-code outlines the whole process which was used to compute reward rate for the different thresholds:

Algorithm 1 Pseudo-code used to generate the reward landscapes

for each $intercept{*}gradient$ combination **do**
  for time (t) values from 1 to tmax **do** ▷ tmax = 50
    $threshold_t = \pm intercept \pm t * gradient$ ▷ threshold height at each time step
  end for
  for trial values from 1 to ntrials **do** ▷ ntrials = 50000
    $time(t) = 0$
    $evidence(e) = 0$ ▷ set initial time and evidence states to 0
    repeat
      $t = t + 1$ ▷ arrival of a new sample
      $e = e \pm 1$ (prob = $0.5 \pm driftrate$) ▷ can support the correct(+)/incorrect(-) option
    until $evidence_t > \pm threshold_t$ OR $t = tmax$ ▷ crossed a threshold/timed-out
    if $e_t > 0$ **then** ▷ correct decision (+)
      $exptime_i = ITI + t * ISI$ ▷ time spent on the trial
      $expscore_i = reward$ ▷ reward for the response
    else if $e_t < 0$ OR $e_{tmax} < threshold_{tmax}$ **then** ▷ incorrect (-)/timed-out decision
      $trialtime_i = ITIp + t * ISI$ ▷ time spent on the trial
      $trialscore_i = penalty$ ▷ penalty for the response
    end if
  end for
  $RR_{intercept*gradient} = \frac{\sum_{i=1}^{ntrials} trialscore_i}{\sum_{i=1}^{ntrials} trialtime_i}$ ▷ compute the reward rate for a given policy
  select the next $intercept * gradient$ combination
end for

2.2.3 Effects of parameter manipulations on the reward landscape

As Table 2.2 shows, there were 6 simulation parameters whose value could be individually (or, if necessary, in combination with other parameters) manipulated to test whether such changes could produce a different degree of symmetry within the reward landscape. Because the manipulation of drift rate values by Malhotra et al. did not affect landscape symmetry as described here (i.e. it changed the optimal policy), this parameter was not examined further, except as a control check that the manipulation of other parameters produces similar changes across different drift rates. Out of the remaining parameters, reward could be ignored, since it uniformly increases the magnitude of reward rates across all policies, without affecting the shape of the landscape.

A potentially interesting set of parameters are the ISI and ITI; on a qualitative level, manipulating each of them individually should have a similar impact on the landscape shape, since both parameters incur additional waiting time between any two decision samples/trials, respectively. More specifically, increasing the value of either parameter would increase the total amount of time during which no reward can be obtained, increasing the value of the denominator in Equation 2.1 and thus reducing the obtainable reward rate. This is why only one of the parameters, the ISI, was manipulated in an initial test of their suitability for changing the landscape symmetry. Specifically, reward landscapes were generated for simulated tasks with ISI = 0.5 and 2, to examine the landscape shape for intervals both smaller and greater than the default value of 1.

As Figure 2.4 shows, the impact of these manipulations was very profound across the entire landscape: The optimal policy for the ISI=2 landscape occurs at a low intercept (1 and 2) and spans a relatively wide range of gradients, leading to a more homogeneous reward rate levels at low intercept and negative gradient values. Intuitively, this should not be surprising - when the rate of stimulus presentation is low, observing more decision samples is more time-consuming. In a context with limited total time, such waiting could prove rather costly, which is why the optimal policy is one that favours quick responses (low intercept, negative gradients). In contrast, the ISI of 0.5 did not look too dissimilar to the default reward landscape, in that a tight cluster of optimal policies can be observed at gradient = 0, albeit occurring at slightly higher intercept values. While interesting in its own right and being successful in changing the landscape symmetry around the optimal policy, this manipulation violated one of the two key conditions: the optimal policy did not remain the same. As such, the ISI and ITI were deemed unsuitable for the current project's purposes.

Thus, two parameters remained - monetary penalty and ITIp. Incidentally, both share a common feature - these parameters affect incorrect responses exclusively, meaning they are intimately linked to accuracy on the task. Examining the 'mean accuracy landscape' for each decision policy of the default simulation (Figure 2.5) revealed a clear and intriguing pattern: a

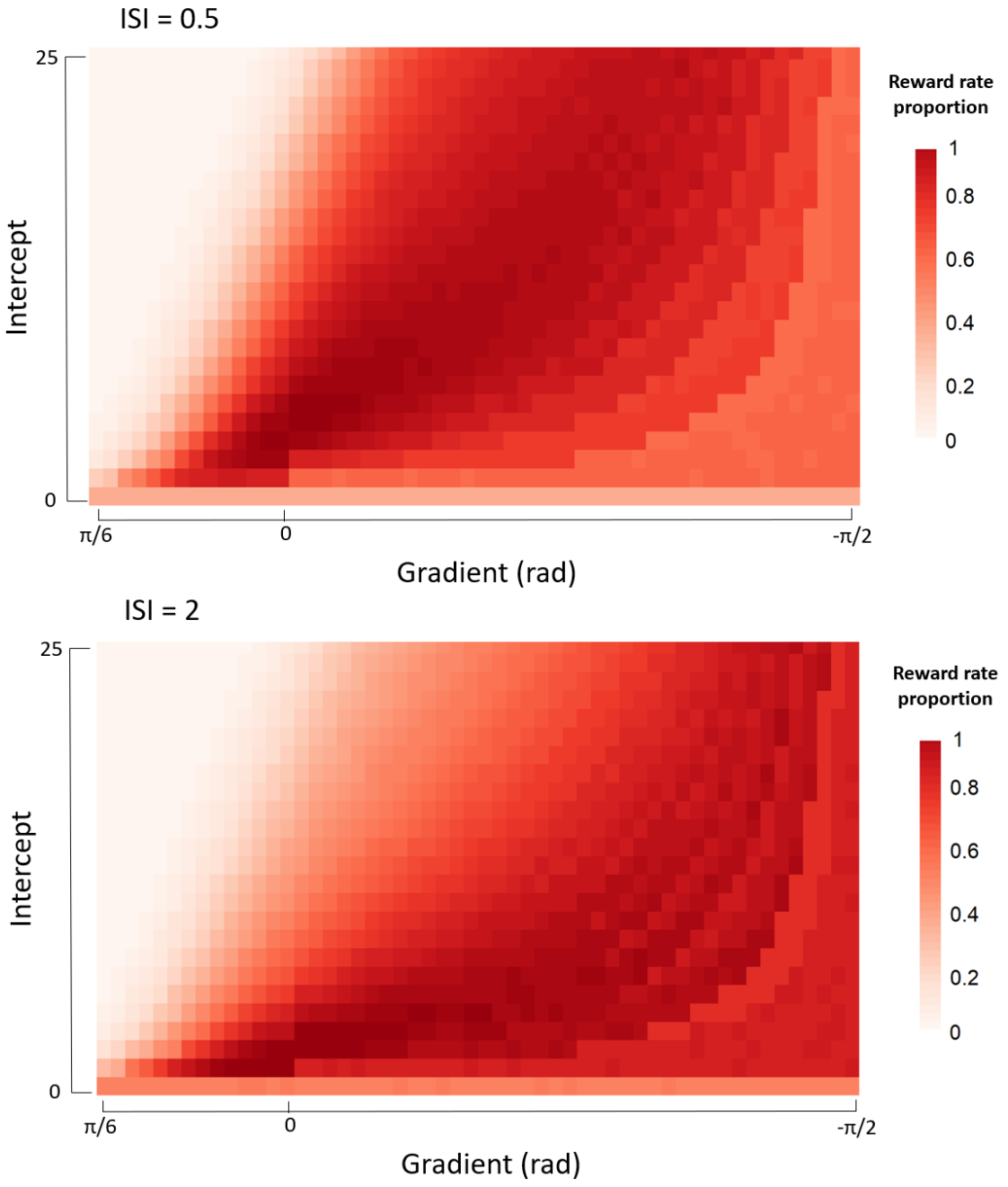

FIGURE 2.4. Comparison of landscapes for a simulation with $ISI = 0.5$ (upper) and $ISI = 2$ (lower). Note that, for visualisation purposes, the scale does not show the absolute reward rate, but instead expresses it as a proportion of the optimal reward rate for a given landscape. The value of optimal reward rate was 0.0466 for the former, 0.0310 for the latter landscape.

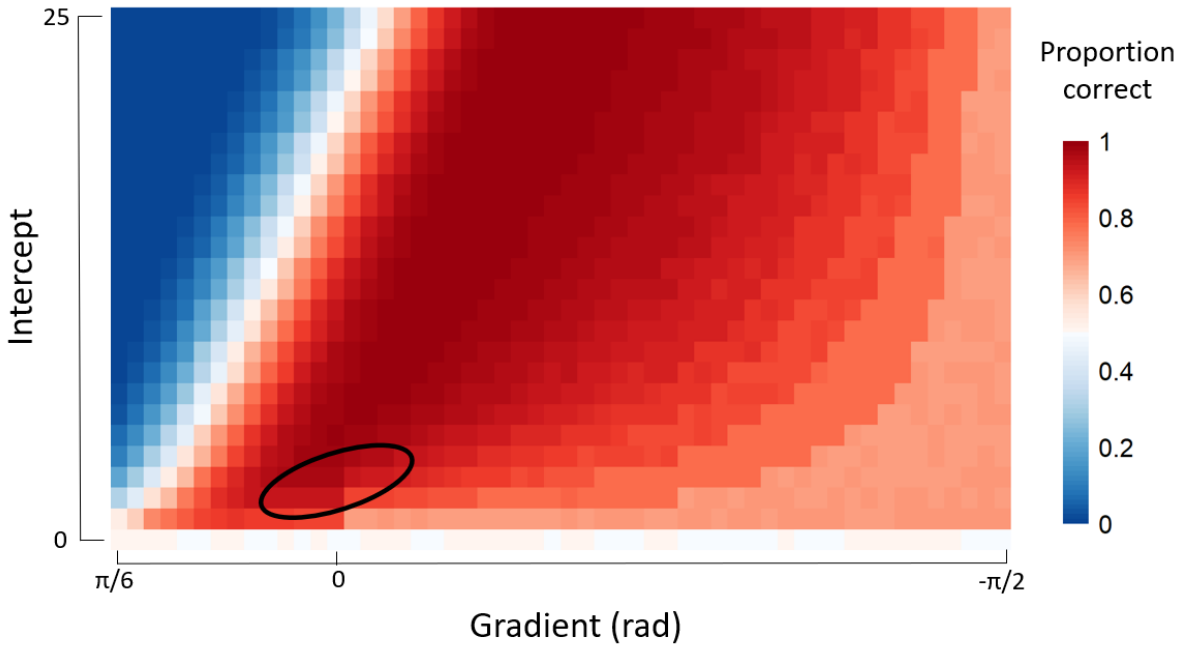

FIGURE 2.5. A heatmap showing the mean accuracy for each decision policy, in the default version of the simulation. Note that below-chance performance (proportion correct < 0.5; blue colour) only occurs when a large proportion of trials time out. That is why only policies with high intercept and/or positive gradient show below-chance performance. The black ellipse shows policies which yielded optimal levels of reward rate.

high percentage of correct choices is made around the optimal policy (black ellipse), with the proportion of these decreasing steadily as one moves towards the more negative gradients. This discrepancy indicates that changing the task parameters which relate to incorrect decisions (monetary penalty, ITIp) should affect reward rates at highly negative gradients profoundly, while the optimal reward rate value should remain largely the same. Given this gradual decrease in reward rate as a function of accuracy, the penalty manipulations seemed like a viable candidate for the simulation's main purpose – to produce a largely one-sided asymmetry in the reward landscape, while keeping the same optimal policy.

To examine the suitability of such manipulations, values of both monetary penalty and ITIp were increased individually and reward landscapes were subsequently generated. For both task parameters, two levels of values were selected and the resulting reward landscapes were compared to the default landscape; for penalty, the values of 1 and 2 were tested (default 0); for ITIp, the initial value of 35 was doubled and tripled, giving the values of 70 and 105. Because the differences are difficult to infer from individual landscapes alone, Figure 2.6 shows a heatmap of

the default landscape being subtracted from the landscapes with the largest penalty value. Although the magnitude of differences varied between the two types of penalty manipulations, both of them confirmed the prediction based on accuracy levels for each policy. That is, an increase in penalty led to a comparatively significant decrease in reward rate at the policies with highly negative gradients, whereas reward rate for the optimal policies (black ellipse) remained largely constant. As such, these difference heatmaps suggest that both penalty manipulations potentially satisfy the two conditions of the target manipulation, as outlined earlier.

In order to examine the degree of asymmetry in these landscapes more closely, Figure 2.7 shows a cross-section across gradients for these landscapes, at the optimal intercept value of 4. In line with the inferences drawn from the accuracy heatmap, both types and both degrees of manipulations kept the optimal policy largely the same - at or very close to the gradient of 0 (although note that the reward rate obtainable from these policies decreased slightly). In addition, both manipulation types resulted in a much more drastic decrease in reward rate for negative gradients, relative to positive ones. Two key differences can also be observed – first, monetary penalty results in a somewhat greater decrease in reward rate than ITIp manipulations, most noticeably for positive and highly negative gradients. Likewise, even though the manipulated parameter values were linearly spaced (i.e. increments of 1 for penalty, 35 for ITIp), changes to the monetary penalty parameter led to a linear decrease of reward rate across levels of manipulation, whereas the magnitude of decrease across ITIp values was non-linear, such that the decrease in reward rate got relatively smaller as the temporal penalty increased. This is largely a consequence of how each parameter is represented in the calculation of reward rate (Equation 2.1) – penalty is found in the numerator, while ITIp features in the denominator, meaning a nonlinear effect is expected.

To compare the precise differences between penalty and ITIp manipulations, cross-sections of reward landscapes for the two manipulations, as well as the default condition, were plotted in the same graph (Figure 2.8). Note that the value of ITIp was deliberately chosen to correspond quantitatively to the $penalty = 1$ condition as closely as possible, with $ITIp = 75$ providing a good approximation. As noted previously, the monetary penalty manipulation yielded somewhat lower reward rate at extreme gradient values than the closely matched ITIp manipulation. For the majority of threshold gradients, however, the levels of reward rate were near-identical across the two penalty manipulations. This is not surprising; reward rate is defined as a ratio of rewards (numerator in Equation 1) and time (denominator in Equation 1). As such, it is possible to keep the ratios similar either by altering the total reward collected, or the total time spent on the task.

As a robustness check, the ITIp and monetary penalty parameter values were manipulated in simulations with a higher (0.3) and lower (0.1) drift rate. Both outputs yielded a pattern

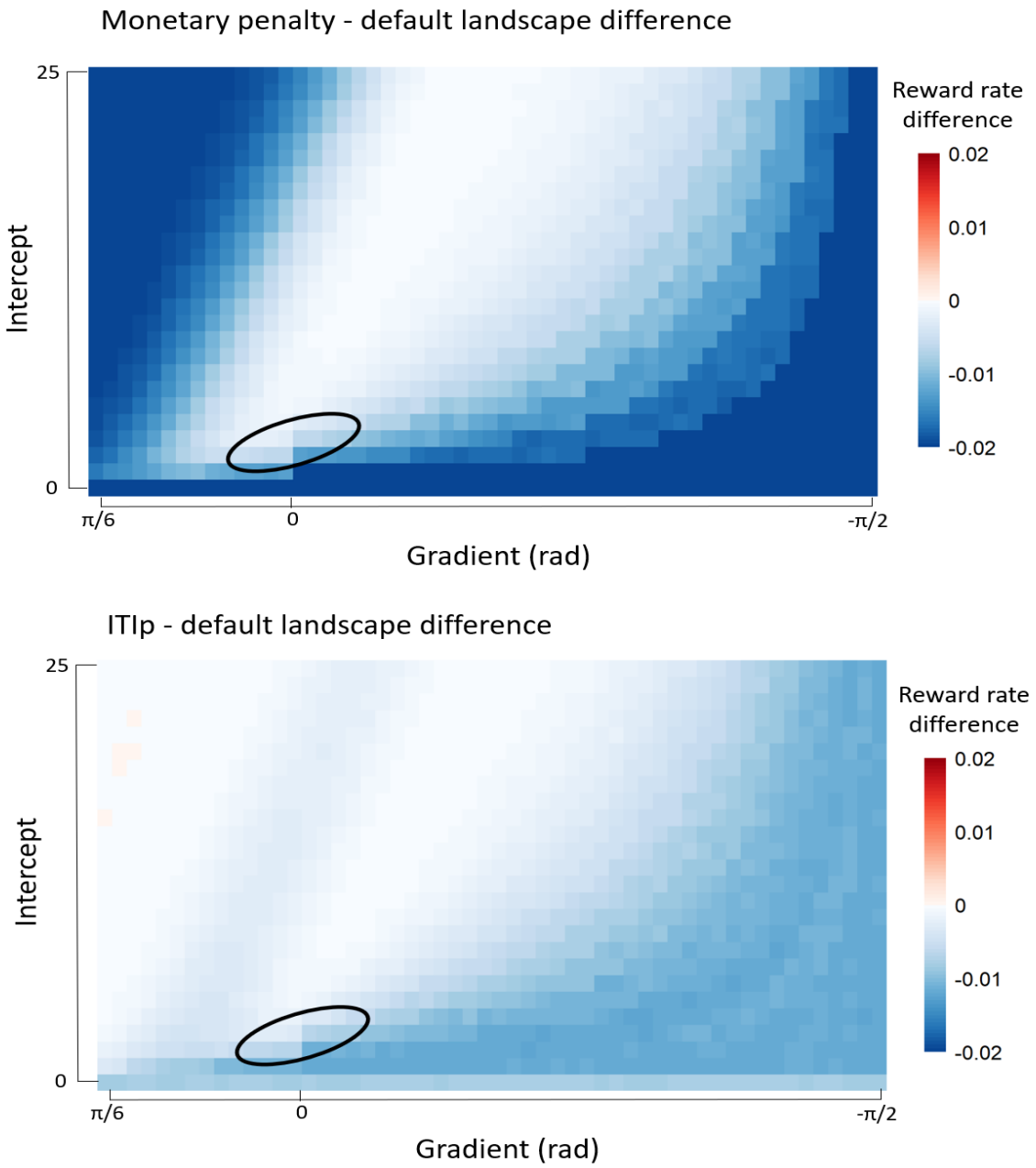

FIGURE 2.6. Heatmaps showcasing the difference in reward rate between the default landscape and the landscapes with increased penalty. The upper panel displays a heatmap where reward rates from the default landscape were subtracted from the landscape where monetary penalty was set to -2; the bottom panel displays a heatmap where reward rates from the default landscape were subtracted from the landscape where ITIp was set to 105. The black ellipse shows the policies which yielded the highest reward rate in the default landscape.

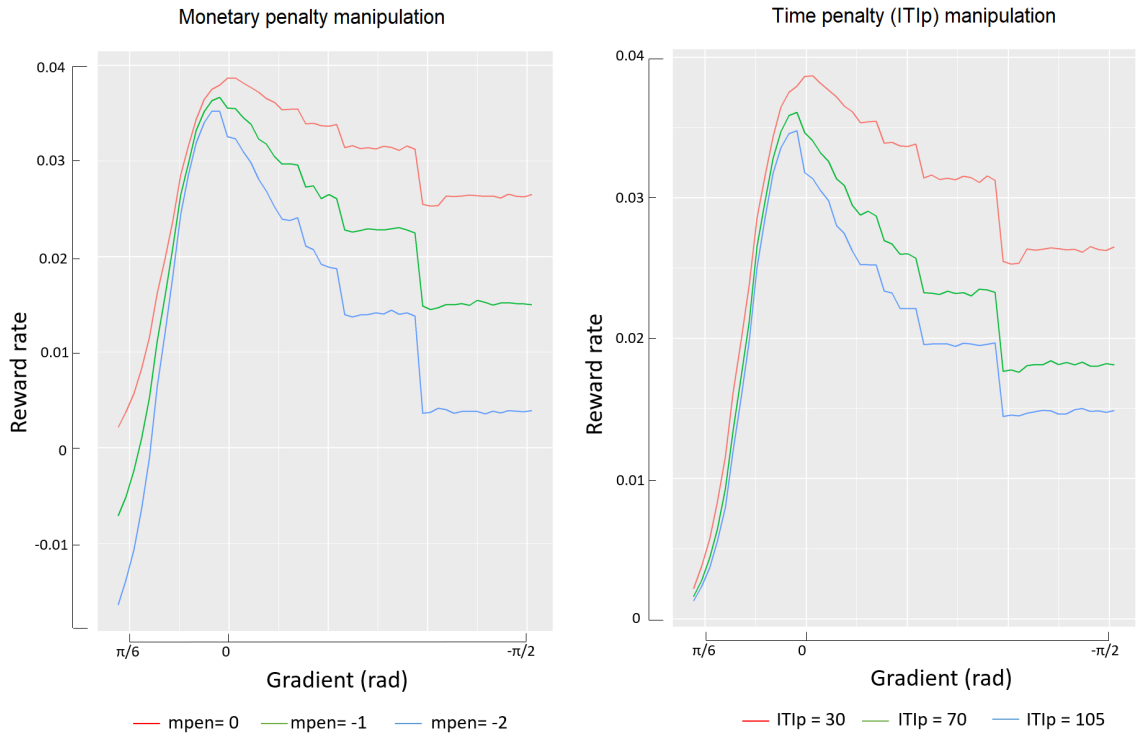

FIGURE 2.7. Reward rates across gradients for penalty and ITIp manipulations. The cross-section were made at the intercept of 3 for both types of manipulations, since this value included the policy with highest reward rate for all conditions

highly similar to the one described in the previous paragraph. This suggests that the effect of penalty-related parameters on landscape symmetry generalizes across decisions with various difficulty levels, and thus it is not simply an artefact of the specific combination of parameter values used in the default version of the simulation.

In conclusion, the penalty-related manipulations produced a landscape where, relative to the default landscape, the same threshold gradients correspond to the optimal policy and the levels of reward rate obtainable by using these policies also remains highly similar. Furthermore, these penalty manipulation landscapes show a noticeable decrease in reward rate levels at policies with negative gradients, whereas reward rate for policies with positive gradients remains nearly identical across manipulations. As such, both monetary penalty and ITIp manipulations satisfy the two conditions outlined in the introduction to this chapter, meaning that the goals of this simulation exercise were successfully met.

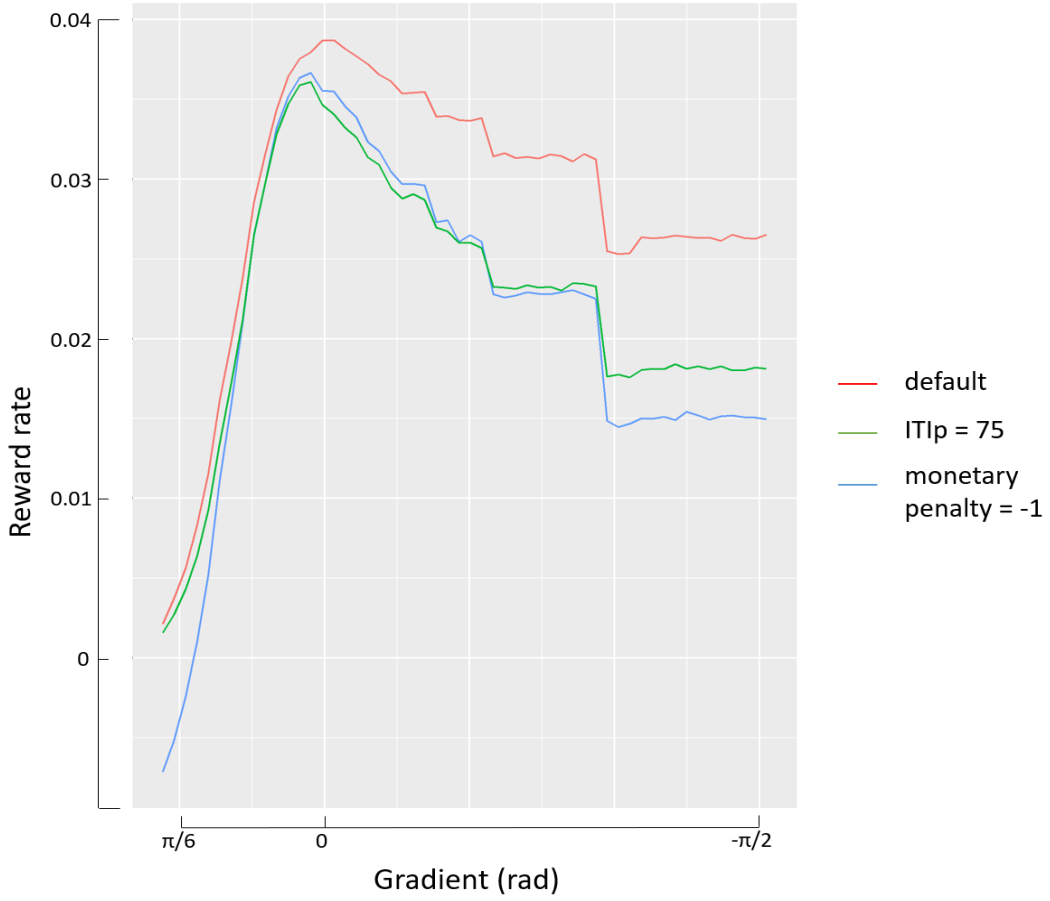

FIGURE 2.8. Cross-sections through the reward landscapes of the time and monetary penalty manipulations, made at the optimal intercept value of 4. Relative to the default landscape, both types of penalty manipulation yield lower reward rate at negative gradients

2.3 Discussion

The aim of the computational work described here was to identify parameters of a simulated decision task, whose manipulation would change the symmetry of the reward rate landscape. This goal was accomplished, with both monetary penalty and time penalty (inter-trial interval for incorrect trials) being capable of increasing the landscape symmetry in highly similar ways. Both of these parameters have a clear counterpart in an experimental setting and they can be easily manipulated in a way that participants are aware of these changes, which makes them ideal for investigating whether decision-makers adapt their choice of policy in response to external factors which affect the reward rate distribution.

2.3.1 Penalty in decision-making

According to the simulations, increased penalties make the policies with negative gradients less beneficial in terms of obtainable reward rate. Thus, an increase in penalty should promote the use of more cautious decision policies - those with a less negative gradient - which is an intuitively plausible notion. The 'loss aversion' effect, where the utility of gaining a certain amount is perceived as lower than the (negative) utility of losing the same amount (Kahneman et al., 2019), has been well-documented in value-based decisions where the potential gains and losses are known in advance (Erev et al., 2008). In addition, penalties with higher magnitude seem to promote a less risky response pattern on the Balloon Analogue Risk Task (Bornovalova et al., 2009), further supporting the notion that increased penalty should promote more cautious decision strategies, which can be achieved by increasing the threshold height and/or selecting a more positive threshold gradient.

However, few empirical investigations of how penalty affects perceptual decision-making have been conducted thus far. One example is a study by Blank et al. (2013), who varied the amount of monetary penalty on a perceptual categorization task while keeping the reward and other parameters constant, and found that higher values of penalty promoted higher decision accuracy. Like many others though, these authors only considered raw accuracy data across penalty manipulations, rather than the deviation from the optimal decision policy which are of primary interest to this project. Indeed, because the optimal decision policy was not computed at all, Blank et al. and other similar studies (e.g. Dambacher et al., 2011) cannot provide direct evidence for or against penalty manipulations increasing the likelihood of selecting a more optimal decision policy, only that participants are more cautious - which, depending on the specific context, need not necessarily be the same. Conversely, studies which showed that higher penalty led to a more optimal response pattern on perceptual tasks (e.g. target detection task; Reckless et al., 2013) only analysed the choice data, rather than considering optimality in terms of speed and accuracy of the decisions. As such, even though higher penalties could lead to the selection of a more optimal policy on a perceptual decision task, direct evidence for this is currently absent from the literature.

2.3.2 Reward-based vs 'optimality' analysis

The reward landscape-based analysis presented here also has more general implications for studies of decision-making and how the degree of optimality is operationalised. As previously noted, the approach adopted by Malhotra et al. and in this project differs rather significantly from that of most previous studies; the standard method is to compute the most optimal policy on a given task and then compare the value of adopted policy parameters (intercept, gradient) against those of the optimal policy. While relatively intuitive, this kind of analysis might prove lacking or outright misleading in certain contexts. As could be seen on the example of Malhotra

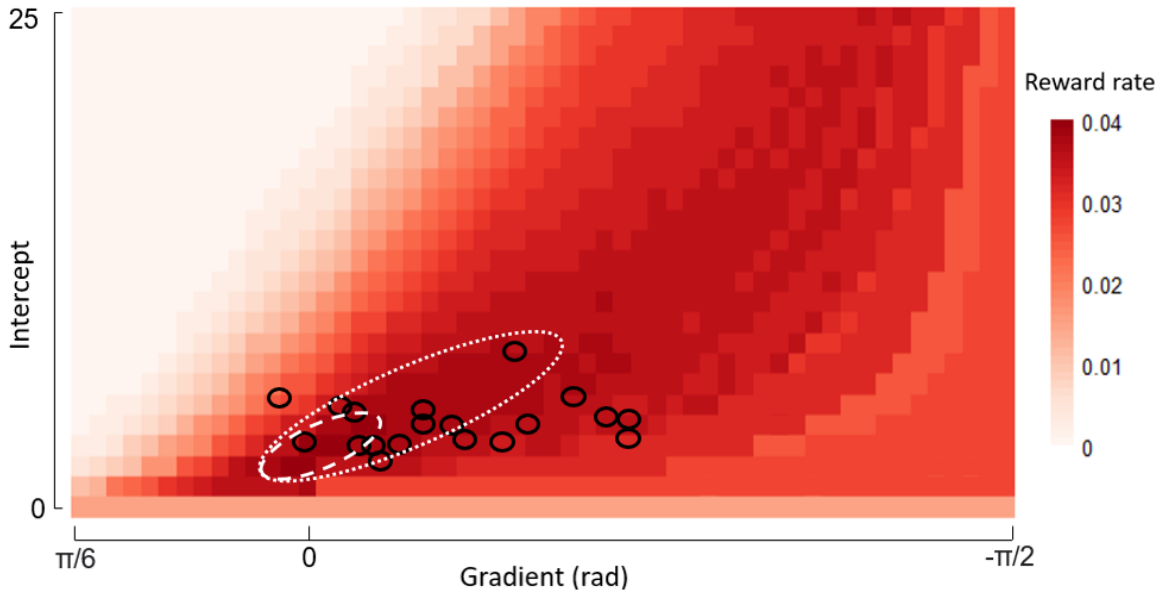

FIGURE 2.9. A reward landscape from Malhotra et al., Experiment 2a, easy decision block (the default landscape). Black dots represent individual-level policies that participants in the study used. Although few individuals selected the policies which yield the highest reward (dashed ellipse), most participants settled in the relatively large region of policies which yield near-optimal reward rate (dotted ellipse).

et al., the optimal policy does not necessarily correspond to a single strategy; instead, a range of decision thresholds can yield optimal or near-optimal levels of reward rate (Figure 2.9). If one only focused on the gradients of adopted policies and compared them to the optimal policy gradient (in this case 0), it could be concluded that individuals displayed clearly sub-optimal decision behaviour due to the numeric difference between threshold gradients. However, calculating the reward rate obtained by the adopted policies shows that, for most participants, they actually yielded near-optimal outcome. Thus, any discussion about optimal decision-making should more explicitly consider what outcome is being optimised (e.g. the reward rate), and how the levels of this outcome are distributed across different decision policies.

Lastly, this reward-based method alone gives valuable insight into the debate about whether individuals employ time-varying decision policies or not. For instance, certain researchers (e.g. Boehm et al., 2020) have suggested that collapsing boundaries are often not utilised by decision-makers because constant policies can provide high enough reward rate, and thus may be seen as the default, 'robust' policy. This assumption holds true for some decision contexts, such as the single-difficulty easy condition landscapes displayed previously. However, by changing certain task parameters, one can easily show that there are contexts where the standard constant

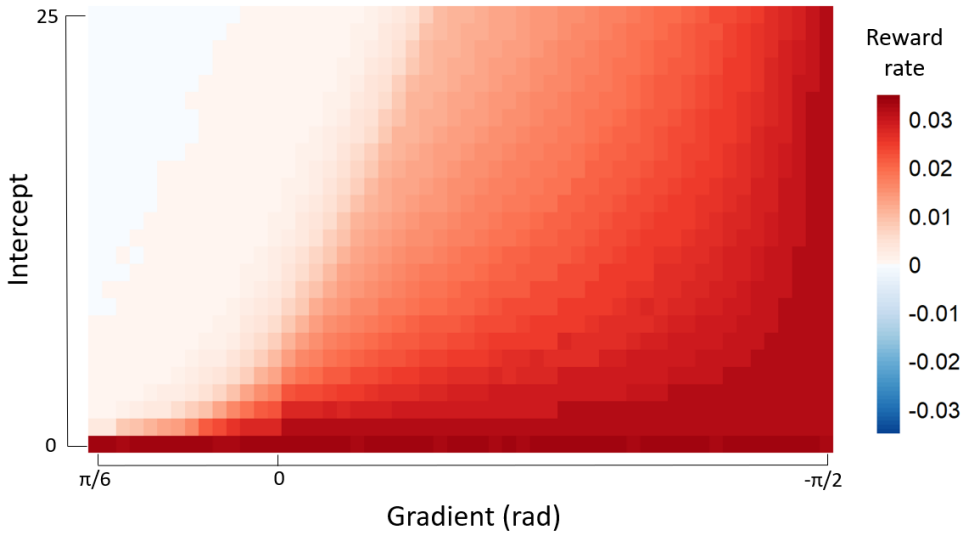

FIGURE 2.10. Reward landscape for a task with drift rate of 0.01 (very high difficulty). The highest reward rate is obtained by guessing - thresholds with intercept of 0 or with highly negative gradient. In contrast, constant policies (gradient = 0 rad) yield very low reward rate.

thresholds are not an optimal or even a near-optimal policy. Figure 2.10 shows one such example, where all parameters are the same as in the default simulation, except that the drift rate is very low (0.01) - similar to the difficult decision blocks used by Malhotra et al. Despite their clear existence, very few empirical studies used contexts where constant thresholds do not represent the optimal or near-optimal policy. Perhaps it is not too surprising - after all, most task parameter manipulations in the current project have not produced landscapes where the optimal policy would deviate too far from zero-gradient thresholds. Nevertheless, the literature on decision policy selection suffers from a 'constant threshold bias', providing grounds for appealing, yet potentially inaccurate over-generalisations such as those made by Boehm et al.

In summary, the computational work succeeded in identifying task parameters whose manipulation changes the reward landscape symmetry on a simulated decision task. These parameters, relating to penalty for incorrect decisions, are both intuitively plausible and easy to implement, making them a viable manipulation in an experimental setting - which will be the focus of the next chapter.

Chapter 3

Testing the effect of reward landscape symmetry on decision threshold selection

3.1 Introduction

The main aim of this project was to test the hypothesis that decision-makers adopt sub-optimal thresholds partly in response to an asymmetry in the reward rate landscape. The computational section of this project established that this degree of landscape asymmetry can be changed by manipulating penalty-related task parameters. To address the research question of interest, these penalty manipulations were now incorporated into an experimental design, to test whether participants respond to changes in landscape symmetry by adjusting their decision policies in accordance with the landscape-based predictions.

3.1.1 Selecting the appropriate manipulation

The simulations described in the previous chapter helped identify two parameters whose manipulation changes the reward landscape symmetry: monetary penalty and time penalty. As the next step, the current section will address the main research question by focusing on one of these manipulations. From the participant's perspective, the effects of monetary penalty might be more easily noticed and interpreted than the effects of time penalty, which affects the value of highest obtainable reward rate rather indirectly - by reducing the total number of trials one can complete, thus reducing the opportunities to accumulate reward. Monetary penalty is also more practical, because decreasing the reward rate with time penalties can easily lead to extremely long waiting times after incorrect responses (>10s). With such long waiting times, there is the risk of reduced participant engagement and poorer data quality (as was the case in Malhotra

et al., 2017, Experiment 2c and 2d). For these reasons, monetary penalty was chosen as the experimental manipulation.

While it would be possible to use the default task parameters from the previous chapter in an experimental setting, note that the default settings include a non-zero time penalty (ITIp = 35). Since the main experimental manipulation involves a penalty manipulation (monetary penalty), the baseline condition should ideally be one without any type of penalty. That is why reward landscapes for a task with no time penalty and different monetary penalty levels were generated and their cross-sections were taken, as shown in Figure 3.1. It is noteworthy that the landscape cross-section for medium penalty condition (monetary penalty = -1; red) shows a reward rate distribution which is very similar to the 'default condition' from Chapter 2, where there was no monetary penalty and the time penalty parameter was set to 35 (7 seconds). Because this landscape is nearly the same as the one in easy decisions from Malhotra et al. (2017), Experiment 2a, we can make the prediction that, much like in Malhotra and colleagues' experiment, participants will use decision thresholds with a negative gradient in the corresponding condition.

The baseline, no penalty condition (monetary penalty = 0; blue) was highly asymmetric in comparison; adopting thresholds with any negative gradient would yield near-optimal levels of reward rate. This means that participants would have little incentive to search for and settle on the policy with gradient of $0rad$ and intercept of 4, since many other thresholds would allow them to achieve the same (or slightly better) reward rate. Consequently, we would expect participants to employ a range of thresholds, with the mean gradient being more negative than in the medium penalty condition. Conversely, increasing the penalty parameter further (monetary penalty = -2; green) increased the landscape symmetry proportionally, making thresholds with negative gradient a significantly less viable strategy. It can be predicted that, in contrast to the no penalty and medium penalty conditions, participants should settle on a policy with less negative gradient in order to retain relatively high levels of reward rate.

3.1.2 Decision inertia

It should be pointed out that the predictions described above hold only under the assumption that individuals would treat the different penalty conditions as completely separate, selecting a new policy every time the context (and, by extension, the reward landscape) changes. However, studies of value-based decision-making often reveal that this may not be the case, because individuals sometimes fail to update their decision preference in the face of change, even when it is no longer beneficial (Alós-Ferrer et al., 2016) - a concept known as decision inertia (Pitz & Reinhold, 1968). In particular, this form of perseverance has been documented in decisions where participants played a time-limited game, in which they had to select between a series of lower but quickly obtainable rewards, and higher but slower to obtain rewards (e.g. Senftleben et al.,

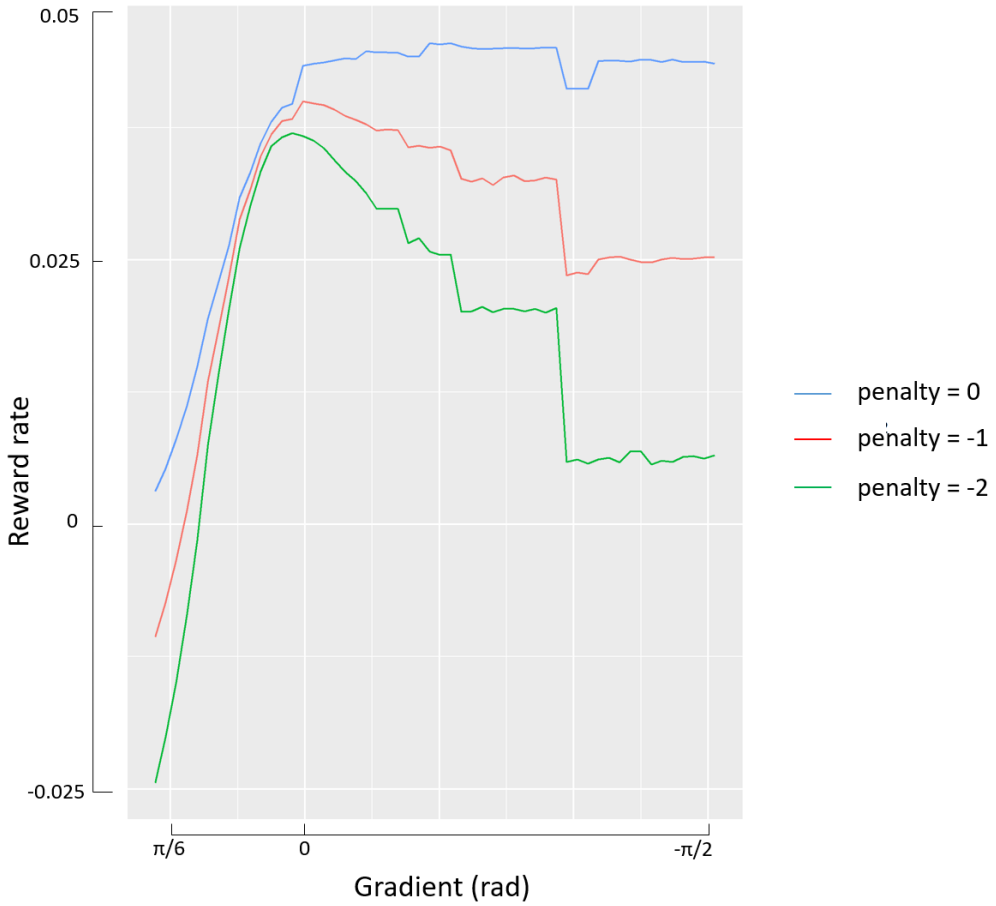

FIGURE 3.1. A cross-section of reward landscapes for a decision task with no time penalty and three levels of monetary penalty. The values of reward rate were calculated at the the optimal threshold intercept of 3.

2019). The reward magnitude of the option initially chosen by the participant was then manipulated on subsequent decisions, such that it gradually increased if the 'low value' option was chosen first or gradually decreased throughout the decision block if the 'high value' option was chosen first. Indeed, participants switched over to the other option at a significantly later point than they should have (i.e. when their preferred option started to yield lower average reward than the other one), suggesting that the initial policy biased their subsequent responses. This is potentially relevant, because participants in the experiments of Senftleben and colleagues were essentially trading off response-related parameters: the value of reward against the time taken to obtain it. Because the expanded judgment task used in Malhotra et al. and in the current experiment is also time limited and involves the trade-off of response-related variables (speed and accuracy across and within trials), it is possible that a similar bias for initially selected decision policies might emerge in this context.

Nonetheless, such an effect has not been tested for in tasks where the evidence is integrated over time and where the key trade-off occurs between speed and accuracy of one's decisions. As a result, it is unknown whether participants would display perseverance for the initially chosen policy in the present task. If present in these tasks, decision inertia would have important consequences in the current study - for instance, participants might select a decision policy in the first decision block they encounter and simply re-use it throughout the experiment, even when the reward landscape changes. In a less extreme example, participants might initially select a policy with negative gradient on a task with highly asymmetric landscape (no penalty), and subsequently be biased towards policies with a negative gradient even in the higher-penalty blocks with a more symmetrical reward landscape, thus obtaining lower levels of reward rate than they would have gained otherwise. As a result, the order in which the different tasks are presented could significantly affect the policies that individuals end up choosing, meaning that the presentation order should be carefully controlled.

Note how the differences in landscape symmetry between monetary penalty conditions depicted in Figure 3.1 offer a relatively straightforward test of the effects of decision inertia: assuming that the choice of decision policy is affected by reward landscape symmetry, then participants who encounter the no penalty condition should select thresholds with a highly negative gradient, whereas those who encounter the high penalty blocks should choose thresholds with gradients close to zero. If decision-makers keep re-using the same decision policy in novel contexts (a key feature of decision inertia), then participants who encountered the no penalty condition first should also use thresholds with more highly negative gradients on the medium penalty condition, compared to participants who encountered the high penalty blocks first. As such, manipulating whether the no penalty or high penalty games are presented first, and subsequently comparing the gradients of thresholds adopted in the medium penalty condition, can serve as a test of whether participants are subject to decision inertia in perceptual decision tasks.

3.1.3 Current experiment

Given the predictions described before, the current experiment has used an expanded judgment task with three levels of monetary penalty, to address the following hypotheses:

1. Participants will use collapsing thresholds in the medium penalty condition, the reward landscape for which is close to the one from Malhotra et al., Experiment 2a (easy blocks). That is, the mean threshold gradient will be lower than 0°.

2. As the reward landscape becomes more symmetrical (higher monetary penalty), the mean gradient of the threshold will be closer to the optimal policy of $\approx 0 rad$. Conversely, with lower symmetry (no monetary penalty), we expect participants to use policies with gradients further from the optimal gradient of 0°.

3. It is possible that the policy used on the first-encountered decision task will bias the selection of decision policy on a subsequent task – an example of decision inertia. If that is the case, then completing the 'medium' blocks after the 'no penalty' blocks should result in a more negative (lower) mean of threshold gradients than when 'medium' penalty is preceded by 'high' penalty decision blocks.

3.2 Methods [1]

3.2.1 Participants

50 participants were recruited through the participant recruitment platform Prolific. All participants received a monetary compensation of £5 for their time, with final earnings partly depending on performance; if the score obtained in the experiment exceeded the threshold of 500p, any additional score was paid out to participants' Prolific account as bonus reward (mean total payment being £6.10).

3.2.2 Experimental paradigm

An important requirement for the current experiment was that it should be directly related to the simulated decision task used in the previous chapter, since the hypotheses are based on the reward landscapes generated by this simulation. As such, we used an expanded judgment task very similar to that used in Malhotra et al. (2017), Experiments 2 and 3. In each trial, participants observed a sequence of up to 50 decision samples (Gabor patches; see *Stimuli and display* subsection), each appearing on the screen for a period of 50ms and then disappearing for 150ms (the period between the onset of two successive stimuli being 200ms). The exact sequence of stimuli was randomly generated at the beginning of each trial, such that any individual sample was likely to be associated with the correct/incorrect response with a probability of 0.7 or 0.3, respectively (drift rate = 0.2). So, if the correct response on a particular trial was 'left', then 70% of samples appeared on the left side of the screen. That said, because participants tended to respond after observing 4-5 decision samples, the actual experienced distribution of samples may differ from the nominal one.

The correct response itself was randomly selected for each decision trial, with 'left' being the correct response on half the trials, and 'right' on the remaining half. Much like in the Malhotra et al. task, each correct response yielded a fixed reward. One key change in the current experiment was that no time penalty was incurred after an incorrect decision. Instead, a fixed amount of score was subtracted after an incorrect decision, the magnitude of which was determined by the current experimental condition. The *Procedure* subsection will describe the procedural aspects of

[1] This experiment has been pre-registered and as such, the methods section is largely adapted from the Open Science Framework pre-registration protocol, which can be accessed through the following link: https://osf.io/8eu5q)

this task in greater detail; nevertheless, Table 3.1 displays all task parameters and their values as employed in the experiment. The task itself was programmed using PsychoPy3 Builder with PsychoJS (Peirce et al., 2019).

Table 3.1: A list of task parameters used in the current experiment, along with their unitless value as specified in the standardised simulation units and the corresponding value expressed in real units.

Parameter name	Value(unitless)	Value(units)
Drift rate	0.2	70%/30%
ISI	1	200ms
ITI	15	3s
ITIp	0	0s
Reward	1	20p
Penalty	0/-1/-2	0p/20p/40p

3.2.3 Stimuli and display

The choice of stimulus was rather arbitrary, since its visual properties should have little effect on decision behaviour as long as it is clearly discernible. As such, we used the same stimuli that were employed in the Malhotra et al. experiments. Specifically, the stimulus used as a decision sample was a circular Gabor patch - a sinusoidal grating viewed through a Gaussian window. The grating had a vertical orientation and a spatial frequency of 1.5 cycles per stimulus width (although see the note in the following paragraph).

Because participants completed the experiment on their own devices, the viewing distance and screen resolution varied between participants and so the stimulus size and frequency could not have been defined in terms of degrees of visual angle. Instead, the stimulus size and positioning on the screen were expressed in terms of PsychoPy's 'height units' (HU), which are defined in relation to the window size: One HU in the 'height' dimension corresponded to the height of the full window, from the bottom edge ($-0.5HU$) to the top edge ($+0.5HU$). The 'width' dimension was scaled in proportion to the aspect ratio. For instance, for the aspect ratio of 4:3, the minimum and maximum values of width would be $-0.5 * 4/3 = -0.667$ (left edge of the window) and $+0.5 * 4/3 = +0.667$ (right edge of the window). This way, setting the width and height dimensions of a stimulus to the same value of HU ensures that the object has the same displayed size in both dimensions. Thus, the size of Gabor patches was set to 0.3 (height) by 0.3 (width) HU. Likewise, the stimulus positions were set to ± 0.3 (height) and ± 0.6 (width) HU. These distances ensured that each stimulus was distinctly positioned in one quadrant of the screen, but not so far from the centre as to be outside the window's boundaries, regardless of the aspect ratio of the participant's window.

3.2.4 Design

The study employed a mixed design. The within-subjects factor was the 'penalty' condition with three levels – 'no penalty', 'medium penalty' and 'high penalty'. In the no penalty condition, incorrect responses incurred no penalty. In the medium penalty condition, incorrect responses incurred a loss of score, its magnitude being equal to the reward for correct decisions (20p). Finally, in the high penalty condition, monetary penalty after an incorrect response was twice as high as the reward (40p).

The between-subjects factor was the order of presentation of these conditions ('order'), with one half of participants experiencing the 'no penalty, medium penalty and high penalty' sequence - 'no penalty first' condition - and the other half completing the 'high penalty, medium penalty and no penalty' sequence - 'high penalty first' condition. Participants were randomly allocated to one of the two order conditions at the beginning of the experiment. Within these conditions, participants completed nine blocks in total (3 blocks back-to-back per penalty condition), in a fixed order; for instance, in the 'no penalty first' order condition, all participants completed three no penalty blocks, then three medium penalty blocks and finally three high penalty blocks.

3.2.5 Procedure

Participants completed the study remotely, on their own devices. They first read a study description outlining the general aims of the study and procedure, which also contained a link to the experiment. Importantly, the description stated that the final payment for participating will be entirely performance-based: points would be accumulated throughout the decision games (separately for each game) and at the end of the experiment, one of these scores would be selected at random and paid out to participant's account. This was to ensure that participants would aim to maximise their score on each of the decision games, by introducing the real possibility of earning low amounts of money if their performance on certain blocks was poor. In reality (unbeknownst to the participants), the minimum earnings were capped at £5 even if the randomly selected score fell below this level, to ensure a fair wage to all participants.

After reading task instructions, participants then completed 10 training trials which familiarised them with the expanded judgment task; no score was gained or lost in these trials and participants were encouraged to try various response strategies. Figure 3.2 provides a schematic visualisation of the sequence of events which took place: each trial began with a 1-second fixation period, consisting of a fixation point in the form of a circle (empty dial) in the centre of the screen. Afterwards participants observed a sequence of flashing Gabor patches (stimulus duration approx. 50ms, with 150ms inter-stimulus interval), with any one sample appearing either on the left or the right side of the screen. At any point during stimulus presentation, participants had to decide whether the stimuli appeared on the left or the right side of the screen

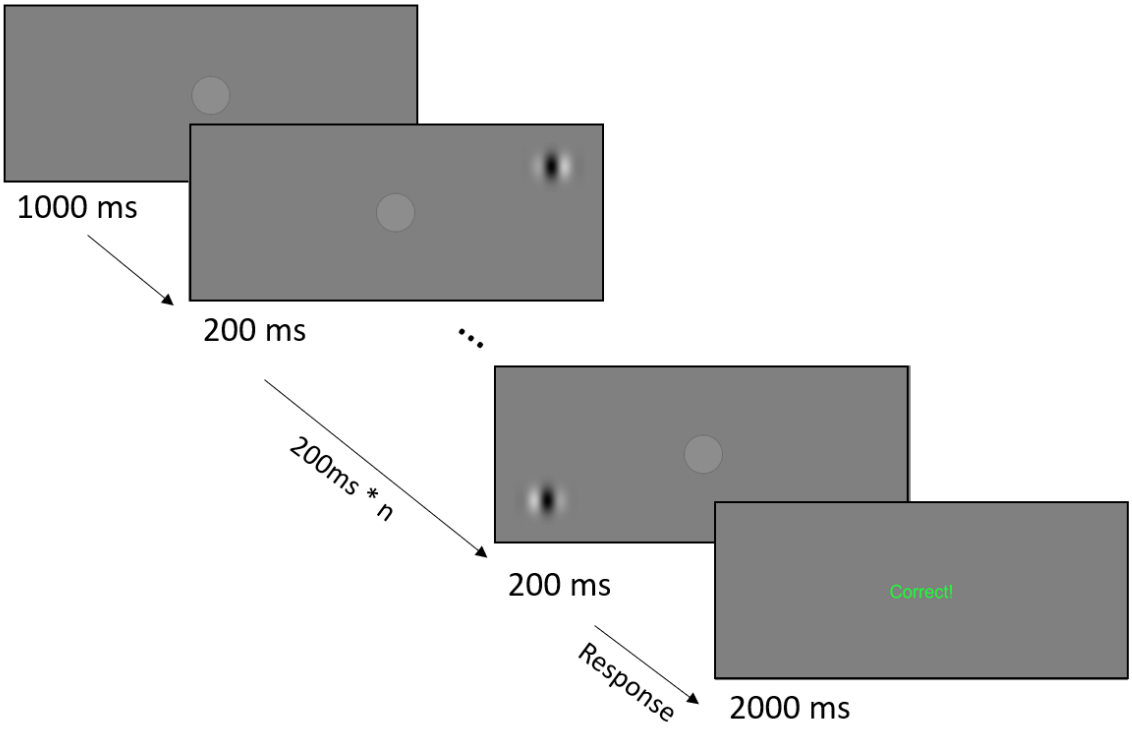

FIGURE 3.2. Schematic visualisation of a single training trial. Each trial started with a 1000ms fixation screen, after which the presentation of evidence samples began. A new sample arrived every 200ms until a response was made or until the trial timed out, at which point the feedback screen appeared for 2000ms and the next trial started.

more often by pressing the 'f' or 'j' key, respectively. If a response was made, or if 10 seconds passed with no input from the participant, the trial was terminated and a feedback screen was displayed for 2 seconds, showing the 'Correct' text in green or 'Incorrect' in red, depending on the response made ('Incorrect' if the trial timed out). The text was then removed and the next trial automatically started, beginning with a 1-second fixation period.

After the training block, the instructions (Appendix B) informed participants that they will play nine 'real' games and outlined the two key differences between these games and the previously encountered training block. First, points would be accumulated throughout each game, such that correct decisions always earned the participant 20p and, depending on the type of the game, incorrect decisions could incur a monetary penalty. At the beginning of each 'game', participants were informed how much penalty will be incurred by an incorrect response in the current block. The current score accumulated in the game so far was displayed during the feedback screen, above the feedback text itself.

Second, instead of completing a fixed number of decision problems, each game was limited to a total of three minutes during which participants could complete as many or few decisions as they wanted. The instructions then emphasised the importance of trading off speed and accuracy of one's responses in order to maximise earnings on each game. To ensure that participants were aware of the remaining time, a dial clock with a dynamic 'arm' was displayed in the centre of the screen at all times, which also served as the central fixation point. At the end of the 3 minutes, the block was immediately terminated and a short, self-paced break started, with text prompting the participants to press the 'spacebar' key whenever they were ready to proceed to the next game.

At the end of the nine blocks, the score from one of the nine games was chosen at random and displayed to the participants (capped at 500p minimum, as mentioned before). Consent was re-confirmed and debrief information was presented, which marked the end of the experiment. In total, the whole procedure took about 40 minutes.

3.2.6 Data analysis plan

3.2.6.1 Variables of interest and transformations

The goal of this experiment was to infer and compare the decision policies that participants used in the different penalty conditions. A constant decision threshold implies that a decision is made solely based on the amount of accumulated evidence. A time-varying decision threshold implies that elapsed time, as well as the accumulated evidence, jointly determine when a decision is made. Therefore, inferring the policy used by participants can be thought of as predicting the 'action' (go/wait) based on the current value of 'time' and 'evidence' variables. The general equation of such logistic regression model can be written as:

$$log(\frac{p_{go}}{p_{wait}}) = \alpha + \beta_x * evidence + \beta_t * time \tag{3.1}$$

with β_x and β_t corresponding to the regression weights for 'evidence' and 'time', respectively, and α denoting the intercept. Solving this equation for $p_{go} = p_{wait}$ reduces the logarithm on the left-hand side to 0, allowing us to find the amount of accumulated evidence for which the probability of 'going' is equal to the probability of 'waiting' expressed as a function of elapsed time:

$$evidence_t = (-\frac{\beta_t}{\beta_x}) * time - \frac{\alpha}{\beta_x} \tag{3.2}$$

This expression gives an equation for a linear function, which can be thought of as a 'line of indifference' (i.e. a set of points where the participant is equally likely to wait and to decide), with gradient $-\frac{\beta_t}{\beta_x}$ and the intercept $-\frac{\alpha}{\beta_x}$. Malhotra et al. (2017) have shown via simulations that the gradient of this line is a good approximation of the gradient of collapsing decision thresholds. As

such, the gradient of this line of indifference was the primary variable of interest in the current study, although the intercept was also computed and compared across conditions, to get a fuller understanding of how the decision strategy changed across conditions.

3.2.6.2 Recorded variables

This informed the choice of variables recorded in the experiment, which were taken at the level of an individual decision sample - i.e. recorded during each stimulus presentation within a decision trial:

- 'action', a binary variable indicating whether the participant made a response when observing the current sample (1) or whether they chose to wait and observe another sample instead of responding (0). This is the dependent variable, which was predicted in a regression model by the variables outlined below.
- 'time', the ordinal number of the current sample in a given decision trial; the first sample had the value of 1, the second 2 etc.
- 'evidence', indicating the objective state of the accumulated evidence after observing the current sample and computed as follows: to mimic the random-walk nature of evidence accumulation process, each observed sample that supported the correct decision was encoded as 1, and samples that supported the incorrect decision took the value of -1. The measure of 'evidence' was a cumulative sum of all sample values observed in the trial. For example, if the correct response on a trial was 'right' and the sequence of stimuli observed so far was 'left' 'right' 'left' 'right' 'right', the 'evidence' variable would be equal to -1 + 1 + (-1) + 1 + 1 = 1 [2]

The more traditional measures of reaction time and accuracy were also recorded, for initial checks of how the penalty manipulations affected the response strategy. Similarly, the scores obtained in each decision block were also recorded, to examine how these response strategies affected the total outcome.

[2]This method of encoding evidence assumes that there were two distinct decision thresholds at which participants could have made their decision - one for correct, the other for incorrect response. In other words, there are actually three action states contained within the data - wait, go(correct) and go(incorrect) - whereas the logistic regression model only accommodates two actions (wait, go). As such, the more appropriate transformation would involve encoding the evidence variable with respect to the decision option selected by the participant, such that evidence samples supporting the selected option would always be encoded as +1, regardless of whether the response was correct on a given trial. In the in-text example, the sequence of observed samples would be encoded as shown if the participant selected 'right'; if they selected 'left', the sign of each individual sample would be changed from positive to negative and vice versa. Because this discrepancy was only noticed after the data has been analysed, a mixed-effects logistic model was fit to two data sets which differed in the encoding method of the evidence variable only, to test whether the extracted regression coefficients would differ substantially. The model predicted highly similar regression weights across both data sets, which is why it was deemed acceptable to report findings of the original analysis, even though it was conducted on the data set with the less appropriate encoding of evidence.

3.2.6.3 Model specification

As described previously, the decision threshold gradients may be estimated using a logistic regression model, using 'time' and 'evidence' as predictors. The regression coefficients can then be transformed into the estimates of intercept and slope of the decision threshold. However, the main point of interest was the comparison of decision thresholds across different penalty conditions. The hierarchical regression framework is ideal for these purposes, since it allows the inclusion of the penalty effects within the same model used to estimate the threshold gradients (as the hierarchical variable), while accounting for the within-subjects element in the data structure. Moreover, performing the hierarchical regression in a Bayesian framework (Shiffrin et al., 2008) enabled the estimation of the population-level distribution for the mean gradient in each penalty condition, as well as the individual-level gradients (obtained by adding specific participant-level residuals to the population-level coefficient estimates).

As a first assessment of whether penalty had an effect on the decision policy, several models were compared. These models allowed for individual-level differences in decision policies, which translate into differences in the intercept and regression coefficients of the logistic regression model. Model 1 was a baseline model that only allowed for the effects of 'time' and 'evidence' to vary across participants — it embodies the assumption that participants vary in their adopted decision policies, but that their own policies remain constant across penalty conditions. Model 2 also allowed the effects of 'time' and 'evidence' to vary with penalty condition, assuming that penalty manipulations affect the selection of decision policy, but that this effect is constant across participants. Model 3 also allowed this effect to vary across participants in a penalty-by-participant interaction, the assumption being that penalty manipulations will have a different effect on different participants. The equations for these three models, complemented by their implementations expressed in brms syntax (Bürkner, 2017), were as follows:

Model 1 (baseline)

$$\pi_{it} = f\left[\alpha + \alpha_{part[i]} + (\beta_t + \beta_{t,part[i]})t + (\beta_x + \beta_{x,part[i]})x\right]$$

$$action \sim 1 + time + evidence + (1 + time + evidence|participant)$$

Model 2

$$\pi_{it} = f\left[\alpha + \alpha_{part[i]} + \alpha_c + (\beta_t + \beta_{t,part[i]} + \beta_{t,c})t + (\beta_x + \beta_{e,part[i]} + \beta_{x,c})x\right]$$

$$action \sim 1 + time + evidence + (1 + time + evidence|penalty + participant)$$

Model 3

$$\pi_{it} = f\left[\alpha + \alpha_c + \alpha_{c:part[i]} + (\beta_t + \beta_{t,c} + \beta_{t,c:part[i]})t + (\beta_x + \beta_{x,c} + \beta_{x,c:part[i]})x\right]$$

$$action \sim 1 + time + evidence + (1 + time + evidence|penalty/participant)$$

In these equations, π_{it} denotes the probability of participant i undertaking the 'go' action at time t, α is the general intercept and β_t and β_x are the regression weights for time and evidence, respectively. Likewise, the subscript x denotes a particular state of evidence, t a particular state of time and c indicates a specific penalty condition. Finally, $f[]$ is the binomial link function.

The models were implemented in the R package 'brms' (Bürkner, 2017). Due to the large size of the dataset and the paucity of information on the recommended prior values of predictors (time, evidence) used in the current models, all regression coefficients had the default prior distributions – weakly informative half-student t-priors on the 'random-effect' coefficients (participant and condition), and uninformative flat priors on the population-level regression coefficients. The models were run with four parallel chains, each one sampling the parameter space over 3500 iterations, with the first 1500 being part of the warm-up routine. As such, the posterior distributions consisted of the total of 8000 draws for each model coefficient.

Even though participants completed training trials, it is likely that exploring the reward landscape and selecting a decision policy takes some time. As such, participants may have used a substantial part of the first block in each penalty condition to find and settle on a stable decision policy that they would subsequently use. That is why these models were fit to two versions of data – one containing the full dataset and one which excluded the first block of each penalty condition. If the inferred policies were similar across these two analyses, it could be concluded that any learning effects occurred rapidly and were not substantial enough to alter the general pattern in decision behaviour. However, there was also the possibility that an effect of penalty

condition would only emerge once the exploration time was accounted for, by excluding the first block from each penalty condition. In that case, the posterior distribution from the entire dataset should differ from that of the 1st-block-excluded dataset.

3.2.6.4 Model comparison and inference criteria

The three models were then compared in terms of how well they fit the data, taking into consideration the differences in model complexity. If the baseline model provided the best fit, it would mean that the penalty manipulation was unsuccessful in producing a change in decision behaviour and that participants likely used the same policy across the different conditions. On the contrary, a better fit of model 2 or 3 would mean that manipulating the penalty promoted a change in decision behaviour. Relative to model 2, the specification of model 3 allows for detecting finer effects in the data – whereas model 2 assumes that penalty manipulation will have the same effect on all participants, model 3 postulates that the penalty manipulation has a different effect on different participants. As before, if this model provided the best fit, it would be assumed that the penalty manipulation leads to the selection of different policies, and that the degree of this effect also varies across participants.

For model comparison, the leave-one-out (LOO) cross-validation method was used - specifically, the PSIS-LOO-CV variant (Vehtari et al., 2017), which is based on the expected log predictive densities (ELPD) of the different models. In essence, this method ranks different models according to how well they predict novel (hypothetical) data, computed from the available dataset. When considering the magnitude of ELPD differences between the different models, the standard conventions were observed: a model was assumed to provide a meaningfully better fit than another if the magnitude of the ELPD difference was several times higher (>3) than the magnitude of the standard error of the ELPD difference.

If either model 2 or 3 (those which included ‘condition’ as a hierarchical level) provided a better fit, the effect of penalty manipulation would be further examined by extracting and comparing the estimated population-level threshold gradients across different penalty conditions. Specifically, these comparisons were performed by assessing the difference between two posterior distributions of mean gradients and the 95% Highest Density Intervals (HDIs) of this resulting difference distribution. If 0 was contained within the HDI, it would be inferred that threshold gradients did not differ between two conditions; if 0 fell outside the HDI, it would be inferred that the thresholds differed between the two conditions (Kruschke, 2014). The same 95% HDI comparison criterion was used when comparing the estimates obtained from fitting models to the full vs. the first-block-removed dataset.

In addition, addressing hypothesis 3 required the comparison of thresholds used in the

'medium penalty' condition, across the two order manipulations. Since the models outlined above were fit to the full data - i.e. collapsed across both order conditions - it was not possible to discern the effects of the order manipulation from these models' regression coefficients. That is why a separate model had to be fit to the medium penalty data only, separately for each of the two order conditions. Given that there was only one penalty condition, we fit the baseline version of the model. Posterior distributions extracted from this model were then compared across order conditions, using the 95% HDI inference criterion outlined in the previous paragraph.

3.2.6.5 Data pre-processing

The inclusion of all decision samples from each trial in the data analysis would imply that participants used all the observed samples to inform their decision, and then made their response instantaneously. In reality, motor execution after the decision has been made takes a non-negligible amount of time, which is why evidence accumulation models typically include a parameter known as non-decision time (Ratcliff, 1978). This parameter estimates the time it takes to execute the motor response, encode the sensory information, or perform any other processes which do not directly relate to accumulating evidence towards the decision threshold. Because the regression models outlined in the previous section capture the decision period only - the time between observing the first sample and the sample which crossed the threshold - it was necessary to account for the non-decision period, particularly between the threshold being crossed and the response being recorded.

To this end, the data from each decision trial were processed, in an attempt to exclude any decision samples which were presented after the putative decision threshold has been crossed. This procedure was based on the findings of Malhotra et al. (2017), who made such calculations for a highly similar task with closely-matched task parameters. In their experiments which adopted the ISI of either 50ms or 200ms, the samples presented in the last 200ms of a trial did not contribute to the decision process for most participants. Because of the similarities in experimental design and the task parameter values in the current study, we applied the same criterion to our data and excluded the last recorded sample from each decision trial, for all participants and in all experimental conditions.

3.2.6.6 Data exclusion criteria

Because guessing could represent a valid strategy in the 'no penalty' condition, no participants were excluded on the basis of overall accuracy rate pooled across all conditions. However, responding randomly would lead to negative earnings in the 'high penalty' condition, which clearly runs counter to the task aims (i.e. to maximize score on each game). As such, the exclusion criterion was based on accuracy rate in the high penalty condition.

Specifically, pilot data from 5 participants were used to infer the mean number of trials completed across the three high penalty blocks ($n = 42 * 3 = 126$). Given this number of trials, a one-tailed binomial test was then used to compute the minimal number of correct trials that would be considered significantly above chance. This threshold occurred at $n_{correct} = 72$, which corresponds to a proportion correct of 0.57. As such, every participant's accuracy in the high penalty blocks was assessed compared against the minimum acceptable level of 0.57 proportion correct. None of the participants' accuracy rate fell below this threshold, and so no data was excluded.

3.3 Results

3.3.1 Descriptive statistics

Before fitting the logistic regression models to the data, descriptive statistics were computed on the measures of reaction time, accuracy and score, as an initial check of whether the experimental manipulations had the expected effects. These statistics were computed on the pre-processed data, where samples presented in the last 200ms of each decision trial were discarded as the non-decision time. Note that the means of all three aforementioned variables were first computed separately for each participant in each condition. These individual means were then used to compute the group-level means for each experimental condition, which is what the means reported in the following paragraphs refer to.

Figure 3.3A shows the reaction time data for each combination of order and penalty conditions. In the 'no penalty first' order condition, the mean response times were 1831ms, 1796ms and 2021ms for the no penalty, medium penalty and high penalty blocks, respectively. For comparison, mean reaction times for the same penalty blocks were 1214ms, 1797ms and 2109ms in the 'high penalty first' order condition. The reaction time distribution across penalty conditions is in line with the expectation that higher penalty would promote a more cautious (and less speedy) response strategy. Notably, reaction times were highly similar across order conditions; an exception were the no penalty blocks, with the mean reaction time being smaller when this block was played last. Such an effect can most likely be attributed to practice effects, combined with the lack of repercussions for an incorrect response encouraging a bolder strategy.

Next, figure 3.3B shows accuracy levels in each condition. Starting with the 'no penalty first' order condition, the mean accuracy was 0.866, 0.879 and 0.886 correct for no penalty, medium penalty and high penalty blocks, indicating only a slight increase in accuracy rate with higher penalty. A similar pattern can be observed in the 'high penalty first' condition, where accuracy rates for the same penalty blocks were 0.830, 0.891 and 0.890 correct. Comparison across the two order conditions reveals two interesting observations: first, accuracy rates in the no penalty blocks were lower when these blocks were encountered last than when they were played first.

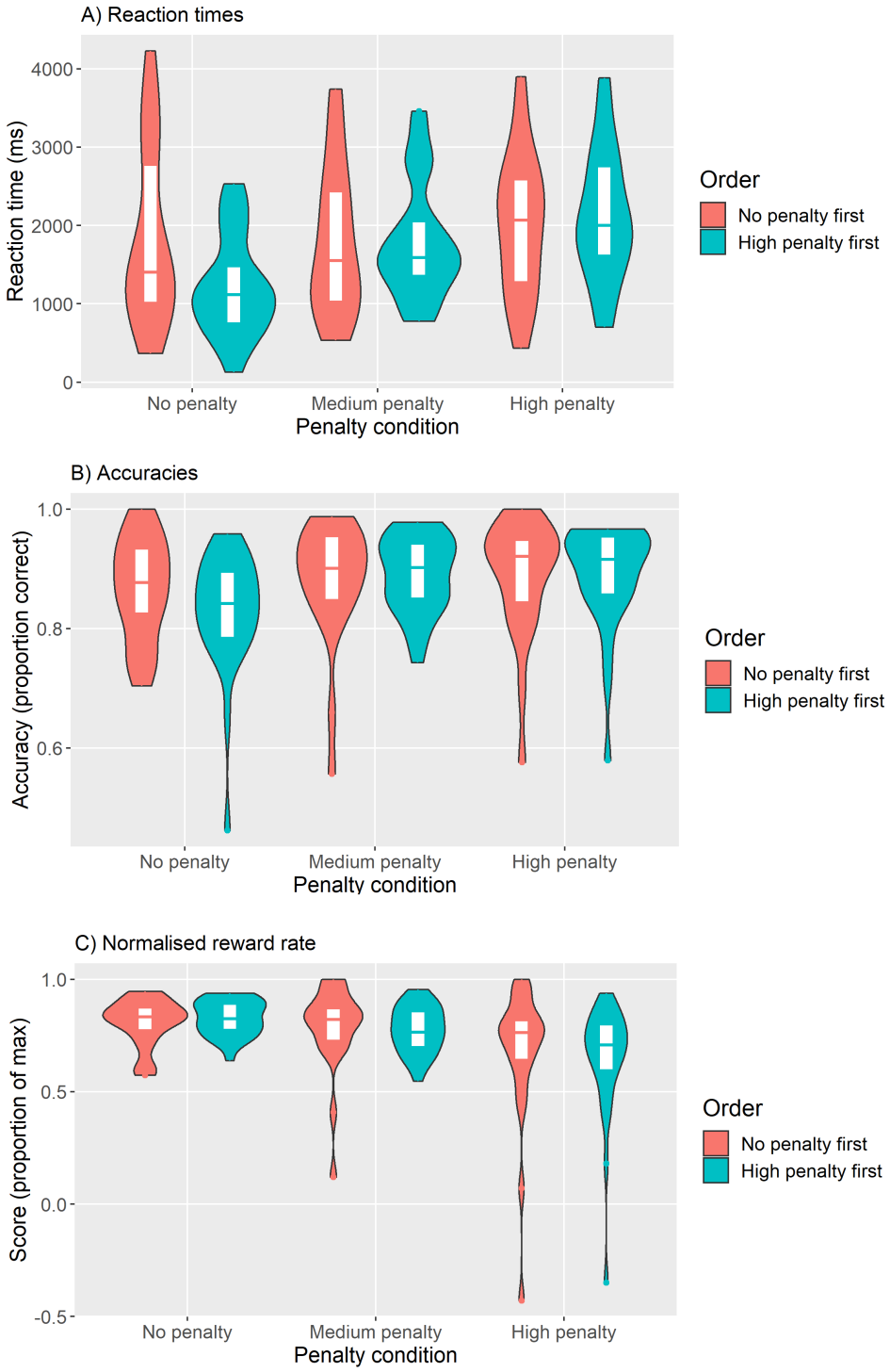

FIGURE 3.3. Violin plots depicting the measures of A) reaction times, B) accuracy rates and C) normalised scores obtained in each penalty condition, across order conditions. Box plots within show the median value, as well as the first and the third quantile.

This relates to the pattern in reaction time distributions where participants responded more quickly in these blocks, further suggesting that a quicker and less cautious response strategy was employed. Second, accuracy rate in the medium penalty blocks was somewhat higher when this block was preceded by high penalty than by no penalty blocks. This might suggest an effect of decision inertia, whereby the more cautious strategy adopted in the high penalty games biased the selection of strategy on the subsequent medium penalty block; however, because participants completed about 120 trials per condition, the accuracy difference of 0.012 translates to making one more correct response, meaning the effect is minute.

The last recorded variable was the score obtained by participants. However, note that, as Figure 3.1 showed, the maximum obtainable reward rate was not identical across the three penalty conditions. To convert these maximum reward rates into the measures of highest obtainable scores, the normalised reward rate that the reward landscapes depict was multiplied by the appropriate task parameter values, much like in Equation 2.3 from Chapter 2. The highest value from the resulting reward landscapes were then extracted, which gave the actual maximum scores of 825p for no penalty, 719p for medium penalty and 668p for high penalty blocks. To extract a normalised measure of the scores obtained by participants, each individual's score was expressed as proportion of the maximum obtainable score in the corresponding penalty condition. As such, these measures are particularly informative, since they indicate how well/poorly participants performed with respect to the optimal policy, allowing for a meaningful comparison between the penalty conditions.

These normalised scores (expressed as proportion of the maximum obtainable score) are shown in Figure 3.3C. In the 'no penalty first' order condition, the means of these scores were 0.807, 0.785 and 0.667 for the no penalty, medium penalty and high penalty blocks. In the 'high penalty first' condition, the scores in the same penalty blocks were 0.828, 0.775 and 0.644, respectively. Regardless of the order condition, the score decreased with higher penalty, suggesting that participants were further away from the point of highest reward rate and, thus, from the optimal policy - contrary to the predictions made by hypothesis 2. Likewise, these measures seem to contradict the hypothesis that individuals would be subject to decision inertia; this account predicts that scores should be higher when the medium penalty condition is preceded by the condition with high landscape symmetry (high penalty), which was not the case. Some order effects were once again present though, with the score on both high and no penalty blocks being somewhat higher when the decision block was encountered last, most likely also attributable to practice effects.

3.3.2 The effect of penalty

The three versions of the Bayesian logistic model were fit to the full data (all penalty conditions across both order conditions) and a LOO comparison was carried out. As table 3.2 shows, Model

3 provided a better fit than the two remaining models. Because the magnitude of the ELPD difference between model 3 and the next-best-fitting model 2 was several times higher than the standard error of this difference, it was concluded that model 3 reliably provided the best account for the observed data. This confirms that the penalty manipulations had an effect on the response pattern, and further suggests that this effect differed across participants.

Table 3.2: Loo comparison of the three model specifications

	ELPD difference	SE of the difference
model 3	0.0	0.0
model 2	-1826.7	63.8
model 1	-2436.9	73.7

To identify how exactly the penalty manipulations affected decision behaviour, population-level posterior distributions for the regression coefficients were extracted from model 3, converted into the measures of intercept and gradient of the line of indifference and subsequently compared across the three penalty conditions.

Figure 3.4 shows the population-level estimates of the intercepts and gradients of the inferred lines of indifference. In the no penalty condition, the mean intercept was 9.27 (SD = 0.80) and the gradient was -16.43°(SD = 2.77°). In the medium penalty condition, the mean intercept was 10.02 (SD = 0.76) and the gradient of -15.63°(SD = 2.50°). Finally, in the high penalty condition, the mean intercept was 11.64 (SD = 0.89) and the gradient of -14.66°(SD = 2.82°). As the figure indicates, the estimates of intercept showed some between-condition variation, but the posterior distributions for the gradient estimates had a large degree of overlap. The inspection of the 95% HDIs of these distributions confirmed that the mean intercepts in all conditions were meaningfully higher than the optimal value of 3 ($HDI_{pen0} = [7.83, 10.93]; HDI_{pen1} = [8.59, 11.49]; HDI_{pen2} = [9.92, 13.42]$). In addition, the 95% HDIs of the population-level gradient distributions did not include zero ($HDI_{pen0} = [-22.43, -11.86]; HDI_{pen1} = [-19.62, -10.46]; HDI_{pen2} = [-20.90, -9.96]$), suggesting that participants consistently used thresholds with negative gradients, much like in the Malhotra et al. experiments.

To establish whether the between-condition differences of these estimates were robust, pairwise differences were taken between each combination of posterior distributions (pen0-pen1, pen1-pen2, pen0-pen2). The 95% HDI's of the resulting 'difference distributions' were then computed; if zero fell outside this interval, then the two original posterior distributions were interpreted to be meaningfully different from one another. As for the posterior distributions of gradients, none of the differences met the 95% HDI criterion ($HDI_{pen0-pen1} = [-7.39, 6.37]; HDI_{pen1-pen2} = [-8.59, 5.53]$; $HDI_{pen0-pen2} = [-9.56, 4.99]$). From intercept comparisons, only the difference between no penalty and high penalty conditions was reliable ($HDI_{pen0-pen2} = [-4.79, -0.27]$), with the other two comparisons failing to reach the same HDI criterion ($HDI_{pen0-pen1} = [-2.81, 1.38]$;

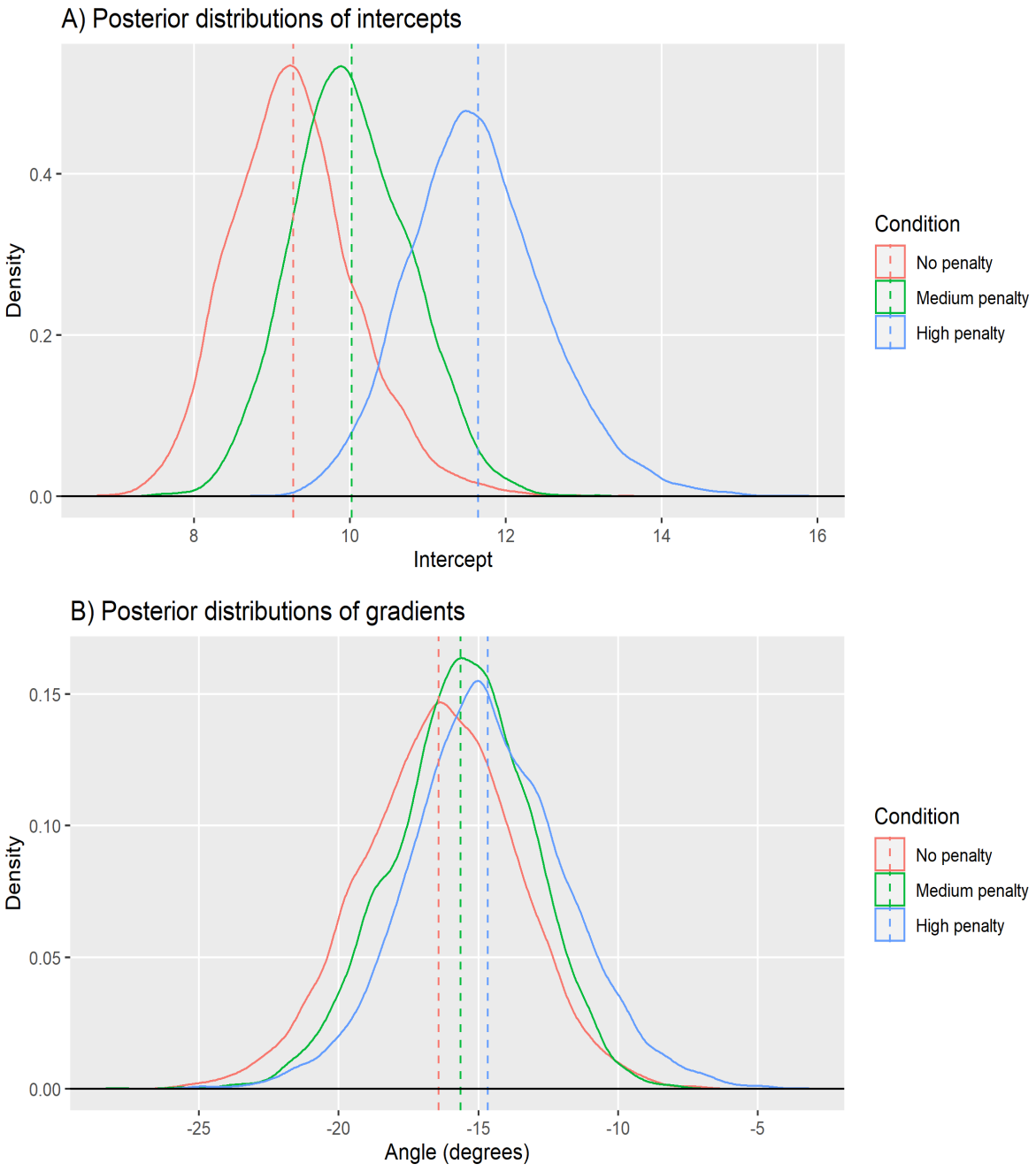

FIGURE 3.4. Population-level posterior estimates of the intercepts and gradients of the line of indifference, across the three penalty conditions. The dashed lines indicate the respective mean values of each density distribution.

$HDI_{pen1-pen2} = [-3.90, 0.54]$). Given this data, the only reliable effect of penalty manipulation was that participants used a higher threshold intercept in the high penalty condition and thus moved further from the optimal policy, whereas the predicted changes in threshold gradients were not observed.

To examine whether the presumed learning period within the first decision block of each penalty condition affected the results, the same comparison was conducted on posterior distributions from model 3, fit to the data where trials from the first block were excluded. The estimated intercepts and gradients of the line of indifference were very similar to the previously reported values: the intercepts were 8.97 (SD = 0.77), 9.89 (SD = 0.74) and 11.48 (SD = 0.95) for the no penalty, medium penalty and high penalty blocks, with the associated gradients of -16.98°(SD = 2.69°), -15.08°(SD = 2.33°) and -15.68°(SD = 2.79°). The 95% HDI inferences led to the same conclusions, whereby gradients did not differ substantially between penalty blocks ($HDI_{pen0-pen1} = [-8.85, 4.08]; HDI_{pen1-pen2} = [-6.34, 6.73]; HDI_{pen0-pen2} = [-8.82, 5.30]$) and the only reliable difference between intercepts was that they were higher in the high penalty than no penalty condition ($HDI_{pen0-pen2} = [-4.93, -0.32]$), with the other two comparisons failing to reach this threshold ($HDI_{pen0-pen1} = [-2.95, 1.15]; HDI_{pen1-pen2} = [-3.90, 0.72]$). Therefore, it can be concluded that learning effects in the first block did not occur at all, occurred very rapidly or conversely, happened gradually over an even longer period of time. As a result, excluding these blocks did not alter the pattern of results: only the intercepts varied between the penalty conditions, but the gradients did not.

3.3.3 Individual-level decision policies

The comparison of posterior distributions suggested some effect of penalty condition on the group-level measures of the line of indifference. However, the differences appear small and one might have expected model 1, which assumed no variance across penalty conditions, to provide an adequate fit to the data. To further elucidate the advantage of model 3, individual-level decision policies (combinations of intercept and gradient) were extracted for the no penalty and high penalty conditions, to determine how individual decision-makers adjusted their decision thresholds between the conditions. Visual inspection of Figure 3.5 indicates that there is evidence of within-participant decision threshold modulation across penalty conditions. However, these effects differ greatly across participants, with some adopting noticeably lower intercepts and near-zero gradients in the no penalty condition, but many more adopting these gradients in the high penalty condition only. One noteworthy pattern is that participants used a somewhat wider range of gradients in the no penalty (SD = 22.42°) than the high penalty blocks (15.34°). Because most negative-gradient policies yielded near-optimal levels of reward rate in the no penalty condition (Figure 3.1), participants could select from a wider range of gradients without losing reward rate and this wider spread might be a direct consequence of this fact. Given the

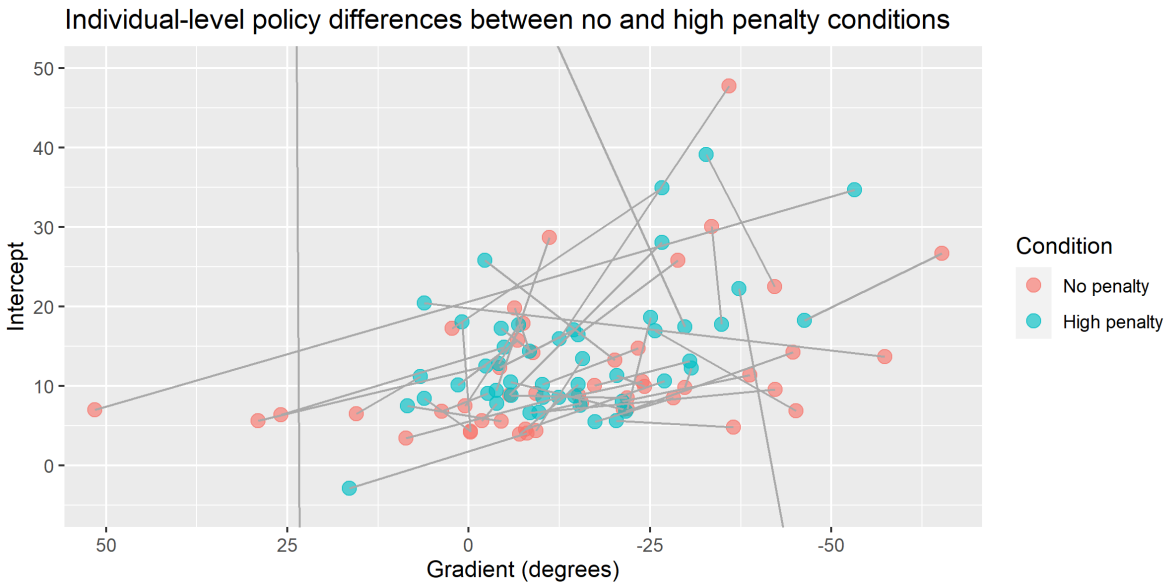

FIGURE 3.5. Individual-level changes in decision policies between no penalty and high penalty conditions. Grey lines connect the policies of the same participant across conditions. The optimal policy was one where intercept = 3 and gradient = 0 for high penalty condition, and intercept = 1 and gradient = 0 for no penalty condition. For visualisation purposes, the limits of the axes were restricted; as a result, there were several (4) estimates which were out of the expected range, as indicated by the solid lines going beyond the axis borders.

above, the population-level inferences should not be interpreted as there being no modulation of decision thresholds across penalty conditions, but rather that these modulations were not systematic across participants and that they 'canceled out' on the group level.

Furthermore, the figure indicates that there were several participants who supposedly adopted very high intercepts (>30) or highly positive gradients (>30°). It is unlikely that these estimates correspond to the actual strategies used by the participants, since they would lead to a much lower (or even negative) reward rate than the participants obtained. To provide representative examples, Figure 3.6 shows a subset of response data from two participants, visualising each individual 'wait' and 'go' action, as well as the model-inferred line of indifference. Note that the logistic regression model was estimating the line of indifference - points where the participant is equally likely to 'wait' and 'go' - which is not equivalent to the decision threshold. Nevertheless, based on prior parameter recovery simulations, the intercepts and gradients of these two types of lines are expected to deviate from one another only slightly. Such was the case for participant 4 (Figure 3.6A), whose inferred line of indifference passes through the bulk of their 'go' data points and generally looks similar to an imaginary line of best fit. In contrast, the model pro-

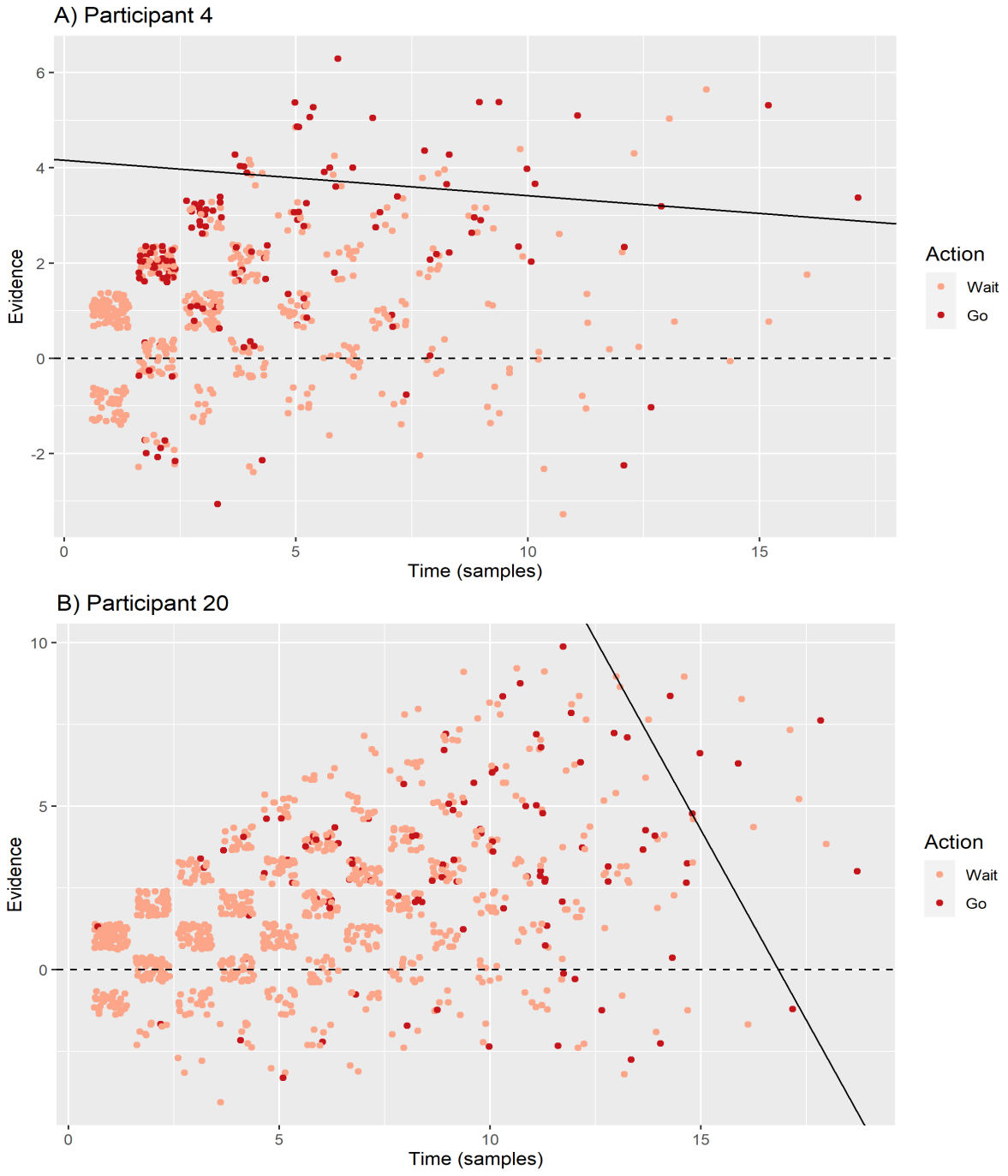

FIGURE 3.6. The action (wait/go) data of two participants, extracted from the medium penalty condition. The dashed line indicates the state of zero evidence and the solid line represents the line of indifference, whose intercept and gradient were inferred by the logistic regression model. For visualisation purposes, the position of individual action points within the graph was offset by a small amount, even though each action could only occur at discrete time/evidence steps. Note that the model provides a relatively accurate description for participant 4 (A), whereas the estimate for participant 20 (B) is biased towards high intercept values.

vided an inaccurate estimate for participant 20 (3.6B) - the line of indifference passes through the tail end, rather than the bulk of data points. That is because the gradient and especially intercept values seem inflated, with intercept being higher than it should be and the gradient also being more highly negative.

A common theme in the group of 'poorly-fitting' participants (n = 5), also visible in the B panel of Figure (3.6), was their inconsistent response pattern: decisions were made at many different levels of time and, in particular, evidence. An informal analysis of individual-level coefficients confirmed that the 'evidence' regression weight was very small (near-zero) for these participants, suggesting that the amount of accumulated evidence had little bearing on when they responded. Because the 'evidence' coefficient is found in the denominator of the intercept and gradient calculations (Equation 3.2), values of both estimates were inflated for these participants. As such, the mapping between the line of indifference and the actual decision threshold of these participants showed a significant bias. This finding indicates that certain individuals' data may be better fit with a different model, one which accounts for less evidence-based response making. Nonetheless, excluding these participants from the analysis only led to a small change in the population-level estimates of intercept and gradient and as such, did not lead to a different interpretation of the current results.

3.3.4 The effect of presentation order

To examine whether the presentation order of penalty conditions had an effect on the selected decision policies, posterior distributions of policy gradients employed in the medium penalty blocks were extracted, separately for each of the two order conditions. As Figure 3.7 shows, the mean population-level gradient in the 'no penalty first' condition was -18.52°(SD = 5.03°) and the mean intercept was 10.48 (SD = 1.72). In the 'high penalty first' condition, the mean gradient was -15.30°(SD = 3.89°), with the mean intercept of 9.93 (SD = 1.49). A 95% HDI analysis confirmed that neither the difference between intercepts ($HDI_{npfirst-hpfirst} = [-7.91, 5.11]$), nor gradients ($HDI_{npfirst-hpfirst} = [-15.84, 9.41]$) was robust.

A similar pattern emerged when only data from the second and third decision blocks were included in the analysis. The mean posterior gradient was -25.71°(SD = 6.28°) with the mean intercept of 10.47 (SD = 1.72) for the 'no penalty first' condition, and gradient of -19.81°(SD = 4.57°) with intercept of 9.93 (SD = 1.49) for the 'high penalty first' order condition. Likewise, the HDI of the differences distributions of gradients ($HDI_{npfirst-hpfirst} = [-20.27, 10.36]$) and intercepts ($HDI_{npfirst-hpfirst} = [-3.91, 5.11]$) contained zero. As such, it can be concluded that the order manipulation has not produced a meaningful change in the gradient of selected policies, regardless of whether learning effects were disregarded or not.

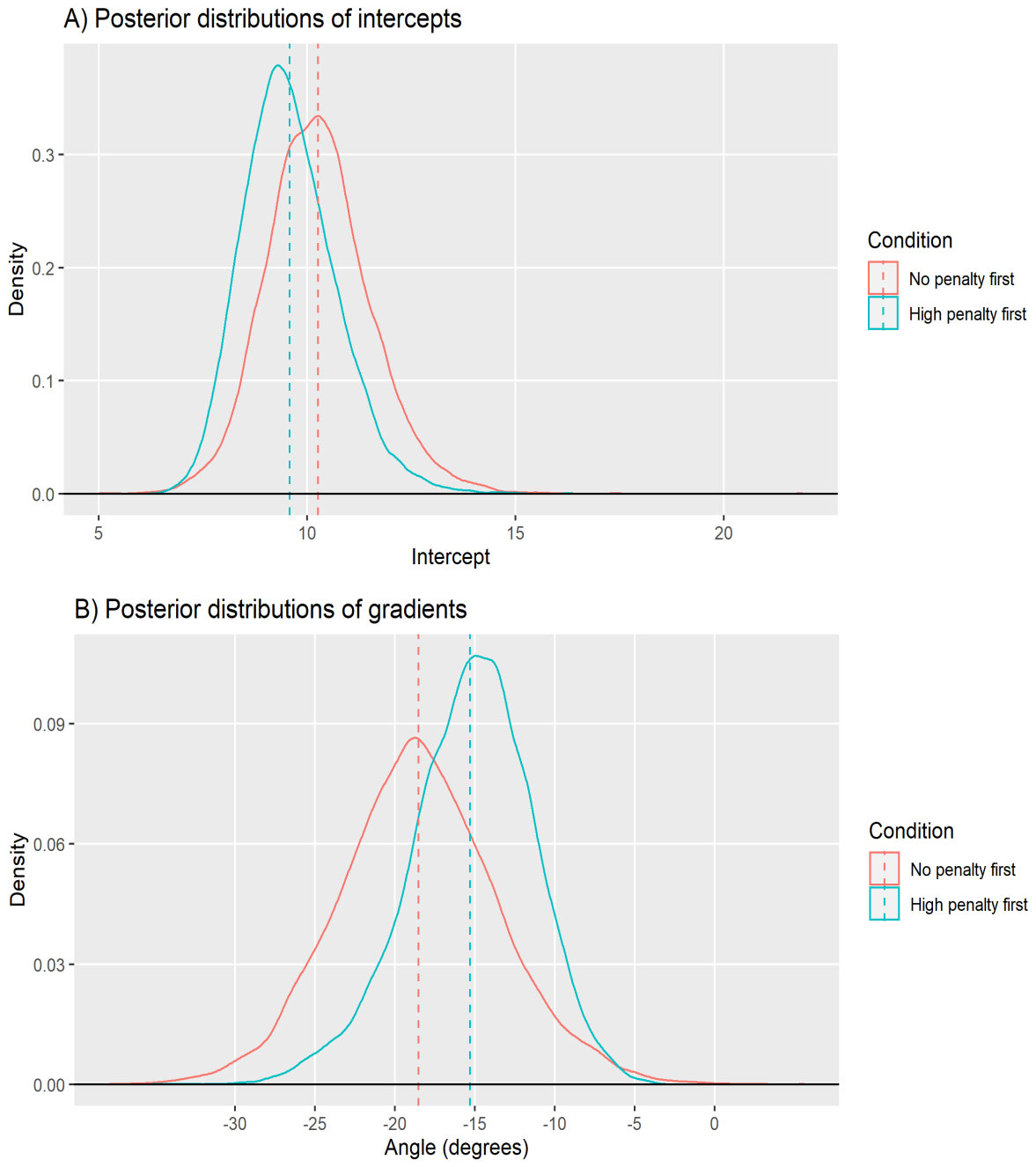

FIGURE 3.7. Population-level posterior estimates of the intercept and gradient of the line of indifference from medium penalty blocks, across the two order conditions. The dashed lines indicate the respective mean values of each density distribution.

3.4 Discussion

This experiment's main aim was to test the hypothesis that the 'shape' of the reward landscape - specifically, the degree of asymmetry in reward rate decrease around the optimal policy - can lead to individuals settling for a sub-optimal decision policy. In line with hypothesis 1, participants used negative-gradient thresholds on the condition with an asymmetric reward landscape, which looked very similar to the reward landscape in Malhotra et al. (2017) and thus replicated their findings. Nevertheless, the prediction of hypothesis 2 - that participants would adjust their threshold gradient as a function of reward landscape symmetry - was not confirmed, as reflected by the lack of substantive group-level changes in decision threshold gradients across penalty conditions. Likewise, there was no evidence that the order in which the penalty conditions were presented had a noticeable effect on the choice of thresholds in a subsequent decision block.

3.4.1 Negative gradient of the line of indifference

As predicted, participants in the medium penalty condition used lines of indifference with gradients similar to those in Malhotra et al. experiments (15°in the current vs. 19°in Malhotra and colleagues' experiment), replicating their findings. This is notable in and of itself - although the present task and analysis were similar to Malhotra et al., the current study used a different experimental manipulation to generate the reward landscapes, which at the very least suggests that the over-collapse is not a response to one specific combination of task parameter values. The fact that the inferred gradients were consistently negative even though the optimal policy was to adopt zero-gradient thresholds is also noteworthy; as mentioned in the General Introduction, certain studies claim that constant thresholds represent the robust default strategy (Voskuilen et al., 2016), or find that individuals do not necessarily use collapsing thresholds even when they represent the optimal policy (Evans, Hawkins, & Brown, 2019). On the contrary, the present results add to the mounting evidence (e.g. Khodadadi et al., 2017, Derosiere et al., 2021) that individuals are capable of employing collapsing thresholds, possibly doing so even when collapsing thresholds do not represent the optimal policy.

However, contrary to hypothesis 2, the mean population-level gradients remained unchanged across the different penalty conditions, regardless of whether the reward landscape symmetry increased or decreased. In a sense, that is not entirely unexpected - the optimal gradient values were about 0°in all penalty conditions and the fact that participants slightly over-collapsed their thresholds, relative to the optimal policy, mirrors the results of most experiments in Malhotra et al. (2017). Nevertheless, the lack of penalty-induced gradient modulation is still peculiar - after all, group-level threshold intercepts were higher in the high penalty condition, meaning that there was some systematic policy modulation across conditions; yet, this modulation did not involve a shift in threshold gradients. At the same time, it is clear that decision-makers are capable of adjusting the gradient of their lines of indifference, such as using a more highly nega-

tive gradient when the optimal policy shifted between the easy-only and mixed-difficulty blocks in the Malhotra et al. experiments. Therefore, a likely explanation for the current results is that a slight over-collapse relative to the optimal policy is the robust default setting in expanded judgment tasks, which is why the reward rate distribution around the optimal policy had little bearing on this effect.

3.4.1.1 Mechanisms of the collapsing thresholds

Although the present analysis lacked the means to elucidate the mechanisms by which participants collapsed their thresholds, it is important to consider why such a peculiar and consistent pattern might have emerged. A likely explanation is that the negative gradients reflected a robust psychological process which was unrelated to strategy selection. For instance, if participants were increasingly more impatient as the trial progressed, they would make decisions at progressively lower evidence levels with more elapsed time - a response pattern that would be captured well by assuming collapsing linear thresholds and one that has been reported in the literature (Boehm et al., 2016). This interpretation could be verified by fitting an appropriate mechanistic model to the data: for example, Hawkins and Heathcote (2019) developed a novel type of evidence accumulation model which does not assume collapsing thresholds, but rather proposes a combination of evidence accumulation process with zero-gradient thresholds and an adjustable temporal deadline, which jointly determine when a decision is made. If such a model demonstrated that individuals set their temporal deadline just before the accumulated evidence would have reached the evidence threshold, while at the same time providing a better fit to data than an EAM with collapsing thresholds, the 'impatience' hypothesis could be a likely cause of the lack of change in gradients.

On a related note, a similar decision pattern would be observed if individuals assigned more value to evidence samples which arrive late in the trial, compared to those early in the trial. Such a notion has been proposed in the form of models that assume an 'urgency signal', which increases with time and results in participants assigning greater weight to samples which arrive late in the trial (Cisek et al., 2009). Note that it can be rather difficult to disentangle this account and the one suggested in the previous paragraph, since the expected behavioural differences are subtle and the main contrast lies in the psychological interpretation of the resulting decision pattern. Nonetheless, Trueblood et al. (2021) recently conducted an extensive model-based analysis of decision-making data, and found that a time-varying urgency signal is compatible with an evidence accumulation framework. As such, the pattern found in the current data could be a result of evidence samples being assigned different importance weights at various points throughout the trial.

One important point to consider is that the apparent collapse of thresholds can occur without the need to alter threshold height within a trial at all. Karşılar et al. (2014) used simulations of

the DDM with various parameter value combinations, and showed that a similar pattern in the data - relatively lower accuracy at longer than shorter reaction times - could be caused via other means; specifically, by using progressively lower zero-gradient threshold across trials, or as a consequence of the between-trial variability in the drift rate. Although the average drift rate remained constant throughout the current experiment, the actual drift rate depended on the sequence of samples generated for the trial and as such, did not remain the same, which could have led to perceived trial-to-trial variations in the drift rate. Malhotra et al. (2017, Appendix C) have shown via simulations that, although the differences between estimated threshold gradients were not significantly affected by high drift rate variability, it led to slight over-estimation of the threshold gradient. As far as trial-to-trial threshold variation is concerned, Figure 3.6 showed a degree of stochasticity in the point at which participants made their response, so it is possible that they did not employ a single set of thresholds across trials. One way to check for this possibility in the future might involve adding 'trial' as a hierarchical variable in the model, or otherwise accounting for possible trial-to-trial threshold variations in the model specification. As a result, the current findings should be taken with some caution, since they may over-estimate the degree to which participants collapsed their decision thresholds.

Regardless of the mechanistic underpinnings of the collapsing thresholds-like decision pattern observed in the current study, it is likely that the negative gradients were not a sign of deliberately chosen decision strategy. Consequently, it is possible that participants were aware of zero-gradient thresholds representing the optimal policy in the task and that those are the thresholds they initially selected. However, the actual decision behaviour ended up deviating from this policy because participants failed to account for impatience or their potential bias when valuating different evidence samples, or the deviation emerged as a result of perceived trial-to-trial differences in task difficulty and participants' response to them. Note, however, that this claim is largely speculative; to test the relative likelihoods of these (and other) explanations, further research which involves a more comprehensive comparison of mechanistic models of decision-making is necessary.

3.4.2 High intercepts and the penalty manipulation

The inferred intercepts of the line of indifference were noticeably higher (≈ 10) than the optimal policy would predict (≈ 4) across all experimental conditions. There is precedent for such an effect in the literature, whereby decision-makers often display an accuracy bias by selecting higher-than-optimal decision thresholds (Zacksenhouse et al., 2010;Balci et al., 2011). Indeed, in one decision-making experiment, only as few as 40% of participants seemed to select threshold height which yielded near-optimal levels of reward rate, with the rest using thresholds which were too high (Bogacz et al., 2010). Subsequent simulations showed that this offset in threshold heights could be accounted for by a (false) assumption that quick responses would lead to poorer

accuracy than they actually would. Interestingly, decision-makers have been shown to set their threshold height more optimally when a response deadline was introduced, which prompted them to respond more quickly than they normally would and thus countered this accuracy bias (Oud et al., 2016). The time-out rate for each decision was rather high in the current experiment - 10s - and so one way of ensuring that participants are closer to the optimal threshold values might be to introduce a shorter time-out rate.

More importantly, increasing the number of training trials has also been shown to reduce the accuracy bias in human decision-makers (Balci et al., 2011). Again, this is not surprising - the training of animal subjects on two-alternative forced choice tasks typically consists of thousands of trials before performance starts to become near-optimal (Stich & Winter, 2006; Mayrhofer et al., 2013). Even human decision-makers usually require extensive training (several hundreds of trials at least) before their behaviour approaches optimality (Evans & Hawkins, 2019). In contrast, participants in the current experiment completed about 120 trials in each of the three penalty conditions, which is many times lower than in the aforementioned studies. As mentioned in the Results section, analysing the data from all three blocks of the three penalty conditions gave nearly identical line of indifference estimates to the analysis of data where the first decision block from each penalty condition had been excluded. While this could indicate very rapid strategy-learning within the first few trials, it is even more likely that significant differences were absent because strategy learning and landscape exploration are an iterative process which spans many hundreds of trials. Therefore, it is possible that participants would have approached near-optimal intercept (and perhaps gradient) values if they were given more extensive training in each penalty condition.

3.4.2.1 The effect of monetary penalty

One potential reason why participants have not approached the optimal policy with higher landscape symmetry is that the monetary penalty was perceived as too punishing. That is, rather than aiming to maximise reward rate by exploring the reward landscape, participants may have focused on avoiding incorrect responses by increasing the threshold height. Higher penalties have been shown to induce a more cautious and less risky behaviour on decision tasks (Dambacher et al., 2011, Blank et al., 2013), an effect quite likely related to the commonly reported psychological aversion to losses and risks in decision-making (Kahneman et al., 2019; Zhang et al., 2014). The current data support this notion, since participants employed higher intercepts in the high penalty condition, even though this shift actually took them further away from the optimal policy and resulted in a lower average reward rate. Thus, it is conceivable that the desire to avoid loss outweighed the urge to test different decision policies, resulting in higher-than-optimal thresholds particularly in the high penalty condition.

Although the previous suggestion to increase the number of trials would likely ameliorate

this issue, the computational section identified another method of testing this proposed explanation - that is, to alter the reward landscape symmetry using time, rather than monetary penalty. Intuitively, the link between monetary penalty and sub-optimal reward rate might be more easily perceived than the less direct relationship between additional waiting time and losing an opportunity to get a potential reward. If this intuition is correct, manipulating the magnitude of time penalty should make participants less prone to being overly-cautious than monetary penalty, prompting them to explore the reward landscape more. The inter-trial interval has received little consideration in the perceptual decision-making literature though (e.g. Evans, Bennett, et al., 2019) and practically none in its potential role as a 'time penalty', which makes it difficult to predict whether this opportunity cost would have a different impact on decision policy selection than a monetary penalty. As such, it would be interesting to compare the psychological effect of the two penalty types in future experiments.

3.4.3 Optimising reward rate

All three hypotheses depended on participants being invested in the task and aiming to maximise score in each decision block. But it is possible that the monetary incentive was not sufficiently motivating and as such, individuals did not actually aim to maximise reward rate. The task conditions were designed in such a way that participants were encouraged to maximise earnings in each decision block: the instructions asked them to do so and explicitly mentioned that balancing the speed and accuracy of their responses is vital; the final payment was (partly) performance-based and it was ensured that each decision contributed to the final earnings by a non-negligible amount; and the final score was randomly selected from one of the nine games, meaning that participants had to try and behave optimally in each block if they wished to maximise their earnings. Consequently, the task design should have been conducive to maximising reward rate. But the study did not involve any attention checks of other tests of participant engagement, and so the possibility that participants were simply not interested in earning as many points as possible cannot be safely ruled out.

One design-related drawback in this regard was that participants were not explicitly notified of the highest possible earnings in each penalty condition, meaning they may have lacked a reference point towards which they should optimise their behaviour. Previous studies show that decision-makers may be able to approach the optimal policy with only basic feedback on the accuracy of each response (Starns & Ratcliff, 2010); but this typically requires extensive training which, as noted above, participants did not receive in the current experiment. Evans and Brown (2017) demonstrated that more detailed feedback can greatly enhance decision-makers' ability to select the optimal strategy on the task; for instance, participants who only got trial-to-trial feedback about whether their last response was correct were slower to approach the optimal decision behaviour than individuals that also received feedback about their overall accuracy

rates and response times in the previous block of trials. Unsurprisingly, the best-performing group was the one that, on top of the trial and block-level feedback, also received guidance at the end of each block on how precisely to adjust their current speed-accuracy balance in order to maximise reward rate and how much more reward the adjustment could earn them. In short, it is possible that the lack of a reference point, as well as the relatively low-level feedback, made the attempts to optimise reward rate in the current experiment challenging and that incorporating more detailed feedback (along with a larger number of trials) could rectify this problem.

3.4.4 Decision inertia

The current study found no discernible effect of the presentation order of penalty blocks on the population-level policy. As described in the introduction to this chapter, the current experimental paradigm shared certain similarities with experiments which identified decision inertia in value-based decisions (e.g. Senftleben et al., 2021). On the other hand, there were also several important differences which could have contributed to the lack of inertia-like effect in this study. For instance, decision perseverance might only manifest when the change in payoff associated with the decision options occurs gradually over time, as in the Senftleben et al. experiments. In contrast, the change in monetary penalty magnitude in the current task happened abruptly between consecutive blocks, and participants were made aware of this change in advance. Therefore, one possibility is that decision inertia is similar to sequential effects observed across areas of cognitive psychology (Yu & Cohen, 2009), whereby the bias only manifests if changes occur on the trial-to-trial level.

A second possibility is that decision inertia is exclusive to decisions where the reward of the presented options is explicitly valuated. That is another major difference between the two experimental paradigms: both decision options in Senftleben et al. offered a reward, as well as a cost associated with obtaining this reward, meaning the preference for one or the other alternative partly depended on the individual's subjective cost-benefit representation of said options. In contrast, decisions in the current experiment included an objectively correct and an incorrect option, only the former of which would yield a (constant) reward. The difficulty of choosing between these two options (the drift rate task parameter) remained constant throughout the experiment and the likelihood of one or the other option being correct on any given trial was also randomly selected. As such, participants in our study had little incentive to pay much attention to whether they are selecting the 'left' or 'right' option on any particular trial and thus were unlikely to associate one option with good outcomes to a greater extent than the other option. In that sense, the current problem of reward landscape exploration and policy selection may be more akin to exploration-exploitation dilemma (Thompson, 1933; Gonzalez & Dutt, 2011), occurring conceptually on a higher and more abstract level, as opposed to making a cost-risk

assessment of two presented options where participants may have a reason to endorse one or the other option.

One last point worth addressing is that the few population-level changes in adopted lines of indifference might at a first glance suggest that decision-makers were subject to an extreme case of inertia - selecting one threshold and reusing it throughout the experiment. However, the plot of individual-level thresholds in Figure 3.5 disconfirms this notion - it is clear that most individuals adapted their decision strategies as a result of the penalty manipulations, albeit not in a systematic manner. In fact, it is the lack of systematic population-level policy changes that makes it difficult to draw any firm conclusions: it could be that an inertia-like order bias simply does not manifest in decision tasks where evidence is integrated over time and where there is no need to valuate the offered options. But it is also possible that such an effect would only arise if there were more discernible population-level changes in the adopted policies to begin with. Therefore, it might prove informative to test for these effects in future work, once the other factors (e.g. higher number of trials) have been accounted for.

3.5 Conclusion

The main goal of this project was to test the hypothesis that the process of decision threshold selection is guided by reward rate distribution throughout the threshold parameter space. More specifically, the project focused on whether the distance between the participant-adopted thresholds and the optimal policy changed as a function of reward rate levels that the sub-optimal policies yielded. The results from an expanded judgment task where monetary penalty manipulations were employed to change the reward landscape did not confirm this hypothesis; people consistently selected sub-optimal threshold intercepts and gradients and, contrary to expectations, failed to adjust their policy in accordance with reward rate levels that the sub-optimal policies yielded. As such, these findings indicate that decision-makers were not systematically exploring the reward landscape before selecting decision thresholds, and that the levels of reward rate given by sub-optimal policies had little bearing on which policy was ultimately chosen.

Nevertheless, there are several important points which prevent the current findings from outright refuting the 'reward rate optimisation' account of decision policy selection. Most notably, participants were subject to the commonly observed accuracy bias, reflected by the high threshold intercepts throughout the experiment but particularly in the high penalty condition. This means that the experimental manipulation used for changing the reward landscape symmetry could have elicited a different psychological response than intended, urging participants to prioritise accuracy preservation over reward rate maximisation. One promising suggestion for future studies would be to increase the number of decision trials per experimental condition, since more extensive training has been shown to override such biases. Such an intervention

might be particularly effective if more detailed guidance and feedback were given to participants, to ensure they are aware of what the optimal policy is and which goal they should aim to optimise their behaviour towards.

Finally, there was evidence that individual participants employed different thresholds across the experimental conditions, but these modulations failed to translate into systematic group-level changes across conditions. This suggests at least some individual-level exploration of the threshold parameter space and as such, it would be premature to conclude that decision-makers are not at all guided by the shape of reward landscape during policy selection, or that they would be unable to approach the optimal behaviour eventually. The present work does, however, confirm the findings of many other studies, that the majority of decision-makers are not successful in maximising reward rate without extensive training and guidance. So, while human decision-makers may still have the potential to act like optimal agents, it seems they are not optimal decision-makers by default.

Appendix A

Comparing the reward landscapes obtained from dynamic programming and from simulation

To check whether the simulation-generated landscape aligned with the landscape obtained by an analytical approach (Markov Decision Process), the landscape generated by the simulation was directly compared to the one from the Malhotra et al. (2017) dynamic programming calculations. Figure A.1B shows this comparison as a subtraction of the 'dynamic programming' landscape from the 'simulation' landscape. As the figure indicates, there were certain differences which could not be attributed to simple noise in the simulation process. Specifically, the simulated landscape seems to predict a larger reward rate at intermediate gradient values, and predict lower reward rate at the 'borderline' policies where a large number of trials times out. A similar pattern emerged when comparing landscapes for difficult($driftrate = 0$), as opposed to easy ($driftrate = 0.2$) decision blocks, suggesting that this is a consistent discrepancy, likely a result stemming from the fact that each method calculates the reward rate in a different way. However, the value of these differences is, on average, two orders of magnitude smaller than the absolute value of reward rates and as such, the simulation-generated reward landscape was deemed accurate enough to proceed with the subsequent manipulations.

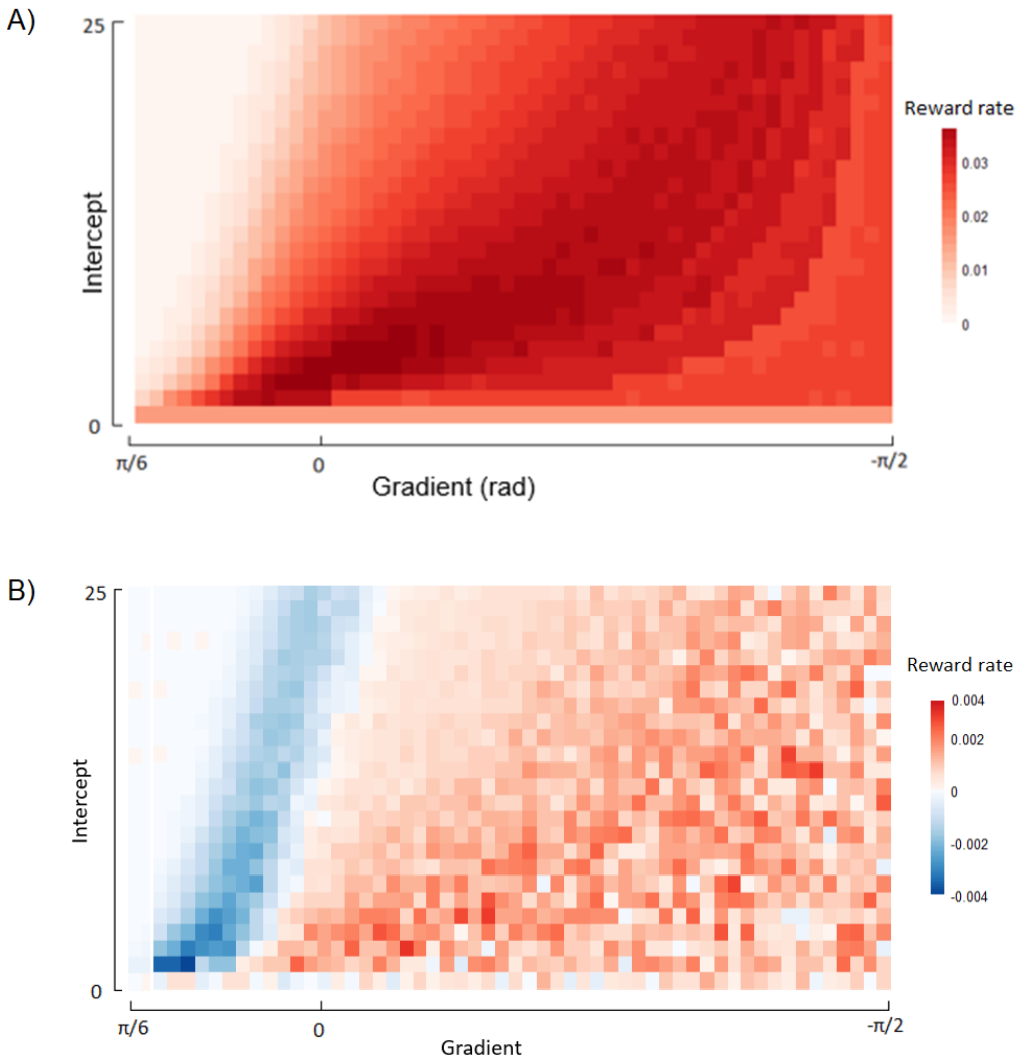

FIGURE A.1. A) The reward landscape generated by a simulation process with 'default' parameter values B) A heatmap of differences, generated by subtracting the 'dynamic programming' landscape from the 'simulation' landscape. Although consistent differences can be observed, note that the scale of this heatmap has been magnified by a factor of one for legibility, meaning that the simulation estimates deviated approximately 1% of the true value of reward rate.

Appendix B

Experimental instructions

This appendix contains verbatim instructions as they were presented to participants at the beginning of the online experiment:

Thank you for your interest in participating! This experiment is based on making a series of simple decisions and it will take about 40 minutes. Please make sure you have read the Prolific study description in full before proceeding. You can press the Esc key to exit the experiment, and simply click the link again once you've familiarised yourself with the study description.

Otherwise, please put your mobile phone on silent and place it out of reach, and try to avoid any distractions. If you are using a laptop, please make sure that it is plugged in and charging. Once you are ready, press 'Spacebar' to go to the consent form.

[CONSENT FORM]

The experiment is designed as nine short games which consist of a series of simple decision problems. In each problem, you will observe a sequence of flashing cues (patterned circle) and decide whether they appear more often on the left or the right side of the screen.

The correct answer for any given problem is always either 'left' or 'right'; so, if the circle flashes more often on the left side of the screen, then the correct answer would be 'left'.

Press 'spacebar' to go to the next screen.

To familiarise yourself with the task, 10 training problems will now follow. You can make your decision by pressing the 'F' key to choose left and the 'J' key to choose right. During the game, always keep the index finger of your left hand on the F key and the index finger of your right hand on the J key. Likewise, always keep your eyes on the circle in the centre of the screen - this will allow you to easily compare the two sides of the screen.

You will receive feedback on whether you were correct after each response, but you will not gain or lose any points. You can respond anytime you want, so feel free to experiment with various response speeds, to find out what works best for you. However, note that if you take too long (>10 seconds), the trial will 'time out' and the next problem will begin automatically.

Press 'spacebar' to start the 10 training problems.

[10 TRAINING TRIALS]

You will now play the nine 'real' games. Although these games and problems are very similar to the training ones, there are also several differences. First, you will accumulate money as you solve problems. If you solve a problem correctly, you will gain 20p. However, you can also lose points, depending on the type of game – incorrect responses can either incur no penalty, result in a loss of 20p or a loss of 40p. You will be notified which type of game you are about to play before it begins.

Keep in mind that the score does not transfer between games, but instead you will get nine separate scores – one for each game. At the end of the experiment, the monetary score from one of these games will be chosen randomly and subsequently paid into your Prolific account. So to maximize your earnings, you should try to maximize the score on each game.

Press 'spacebar' to go to the next screen.

Second, rather than solving a fixed number of problems, each game will have a time limit of 3 minutes. That is why balancing the speed and accuracy of your responses is extra important - if you guess without seeing any cues, you will have a 50% chance of responding correctly. If you wait and see more cues, the chances of responding correctly will increase. However, if you spend too much time on a single problem, you will be able to solve fewer problems over the 3 minutes and collect less money.

A clock arm in the centre of the screen will indicate the remaining time throughout each game.

Press 'spacebar' to go to the next screen.

To summarise: in each game, you will have 3 minutes to solve decision problems and should aim to maximise score by balancing the speed and accuracy of your responses. Each correct response will earn you 20p and depending on the type of game, incorrect response can incur a penalty of 0p, 20p or 40p.

Throughout the game, keep your eyes fixated on the clock in the middle of the screen, with your left index finger on the 'f' key and your right index finger on the 'j' key. Please note that

the response keys might not always register on the first press – this is a known glitch, simply press the same key immediately if this happens. Please keep your attention totally focused on the game for its 3 minute duration. In between games you can take a break and relax.

Press 'spacebar' to start the first game.

Bibliography

Alós-Ferrer, C., Hügelschäfer, S., & Li, J.
(2016).
Inertia and Decision Making.
Frontiers in Psychology, *7*.
https://doi.org/10.3389/fpsyg.2016.00169

Balci, F., Simen, P., Niyogi, R., Saxe, A., Hughes, J. A., Holmes, P., & Cohen, J. D.
(2011).
Acquisition of decision making criteria: Reward rate ultimately beats accuracy.
Attention, Perception, and Psychophysics, *73*(2).
https://doi.org/10.3758/s13414-010-0049-7

Blank, H., Biele, G., Heekeren, H. R., & Philiastides, M. G.
(2013).
Temporal characteristics of the influence of punishment on perceptual decision making in the human brain.
Journal of Neuroscience, *33*(9).
https://doi.org/10.1523/JNEUROSCI.4151-12.2013

Boehm, U., Hawkins, G. E., Brown, S., van Rijn, H., & Wagenmakers, E. J.
(2016).
Of monkeys and men: Impatience in perceptual decision-making.
https://doi.org/10.3758/s13423-015-0958-5

Boehm, U., van Maanen, L., Evans, N. J., Brown, S. D., & Wagenmakers, E. J.
(2020).
A theoretical analysis of the reward rate optimality of collapsing decision criteria.
Attention, Perception, and Psychophysics, *82*(3).
https://doi.org/10.3758/s13414-019-01806-4

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D.
(2006).
The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks.

Psychological Review, *113*(4).
https://doi.org/10.1037/0033-295X.113.4.700

Bogacz, R., Hu, P. T., Holmes, P. J., & Cohen, J. D.
(2010).
Do humans produce the speed-accuracy trade-off that maximizes reward rate?
Quarterly Journal of Experimental Psychology, *63*(5).
https://doi.org/10.1080/17470210903091643

Bornovalova, M. A., Cashman-Rolls, A., O'Donnell, J. M., Ettinger, K., Richards, J. B., deWit, H., & Lejuez, C. W.
(2009).
Risk taking differences on a behavioral task as a function of potential reward/loss magnitude and individual differences in impulsivity and sensation seeking.
Pharmacology Biochemistry and Behavior, *93*(3).
https://doi.org/10.1016/j.pbb.2008.10.023

Brown, S. D., & Heathcote, A.
(2008).
The simplest complete model of choice response time: Linear ballistic accumulation.
Cognitive Psychology, *57*(3).
https://doi.org/10.1016/j.cogpsych.2007.12.002

Bürkner, P. C.
(2017).
brms: An R package for Bayesian multilevel models using Stan.
Journal of Statistical Software, *80*.
https://doi.org/10.18637/jss.v080.i01

Busemeyer, J. R., & Rapoport, A.
(1988).
Psychological models of deferred decision making.
Journal of Mathematical Psychology, *32*(2).
https://doi.org/10.1016/0022-2496(88)90042-9

Cisek, P., Puskas, G. A., & El-Murr, S.
(2009).
Decisions in changing conditions: The urgency-gating model.
Journal of Neuroscience, *29*(37).
https://doi.org/10.1523/JNEUROSCI.1844-09.2009

Dambacher, M., Hübner, R., & Schlösser, J.
(2011).
Monetary incentives in speeded perceptual decision: effects of penalizing errors versus slow responses.

Frontiers in Psychology, *2*(SEP).
https://doi.org/10.3389/fpsyg.2011.00248

de Bruyn, B., & Orban, G. A.
(1988).
Human velocity and direction discrimination measured with random dot patterns.
Vision Research, *28*(12).
https://doi.org/10.1016/0042-6989(88)90064-8

Derosiere, G., Thura, D., Cisek, P., & Duque, J.
(2021).
Trading accuracy for speed over the course of a decision.
Journal of Neurophysiology, *20*(7).
https://doi.org/10.1152/jn.00038.2021

Ditterich, J.
(2006).
Stochastic models of decisions about motion direction: Behavior and physiology.
Neural Networks, *19*(8).
https://doi.org/10.1016/j.neunet.2006.05.042

Drugowitsch, J., Deangelis, G. C., Angelaki, D. E., & Pouget, A.
(2015).
Tuning the speed-accuracy trade-off to maximize reward rate in multisensory decision-making.
eLife, *4*(JUNE2015).
https://doi.org/10.7554/eLife.06678

Drugowitsch, J., Moreno-Bote, R. N., Churchland, A. K., Shadlen, M. N., & Pouget, A.
(2012).
The cost of accumulating evidence in perceptual decision making.
Journal of Neuroscience, *32*(11).
https://doi.org/10.1523/JNEUROSCI.4010-11.2012

Erev, I., Ert, E., & Yechiam, E.
(2008).
Loss aversion, diminishing Sensitivity, and the effect of experience on repeated decisions.
Journal of Behavioral Decision Making, *21*(5).
https://doi.org/10.1002/bdm.602

Evans, N. J., Bennett, A. J., & Brown, S. D.
(2019).
Optimal or not; depends on the task.
Psychonomic Bulletin and Review, *26*(3).
https://doi.org/10.3758/s13423-018-1536-4

Evans, N. J., & Brown, S. D.
(2017).
People adopt optimal policies in simple decision-making, after practice and guidance.
Psychonomic Bulletin and Review, *24*(2).
https://doi.org/10.3758/s13423-016-1135-1

Evans, N. J., & Hawkins, G. E.
(2019).
When humans behave like monkeys: Feedback delays and extensive practice increase the efficiency of speeded decisions.
Cognition, *184*.
https://doi.org/10.1016/j.cognition.2018.11.014

Evans, N. J., Hawkins, G. E., & Brown, S. D.
(2019).
The Role of Passing Time in Decision-Making.
Journal of Experimental Psychology: Learning Memory and Cognition.
https://doi.org/10.1037/xlm0000725

Franks, N. R., Dornhaus, A., Fitzsimmons, J. P., & Stevens, M.
(2003).
Speed versus accuracy in collective decision making.
Proceedings of the Royal Society B: Biological Sciences, *270*(1532).
https://doi.org/10.1098/rspb.2003.2527

Frazier, P. I., & Yu, A. J.
(2008).
Sequential hypothesis testing under stochastic deadlines.
Advances in Neural Information Processing Systems,
1–8.

Gonzalez, C., & Dutt, V.
(2011).
Instance-Based Learning: Integrating Sampling and Repeated Decisions From Experience.
Psychological Review, *118*(4).
https://doi.org/10.1037/a0024558

Hawkins, G. E., Brown, S. D., Steyvers, M., & Wagenmakers, E. J.
(2012).
An optimal adjustment procedure to minimize experiment time in decisions with multiple alternatives.
Psychonomic Bulletin and Review, *19*(2).
https://doi.org/10.3758/s13423-012-0216-z

Hawkins, G. E., Forstmann, B. U., Wagenmakers, E. J., Ratcliff, R., & Brown, S. D.
(2015).
Revisiting the evidence for collapsing boundaries and urgency signals in perceptual decision-making.
Journal of Neuroscience, *35*(6).
https://doi.org/10.1523/JNEUROSCI.2410-14.2015

Hawkins, G. E., & Heathcote, A.
(2019).
Racing Against The Clock: Evidence-Based Vs. Time-Based Decisions.
https://doi.org/10.31234/osf.io/m4uh7

Heitz, R. P.
(2014).
The speed-accuracy tradeoff: History, physiology, methodology, and behavior.
https://doi.org/10.3389/fnins.2014.00150

Herz, D. M., Tan, H., Brittain, J. S., Fischer, P., Cheeran, B., Green, A. L., Fitzgerald, J., Aziz, T. Z., Ashkan, K., Little, S., Foltynie, T., Limousin, P., Zrinzo, L., Bogacz, R., & Brown, P.
(2017).
Distinct mechanisms mediate speed-accuracy adjustments in cortico-subthalamic networks.
eLife, *6*.
https://doi.org/10.7554/eLife.21481

Huber, O., Huber, O. W., & Bär, A. S.
(2014).
Framing of decisions: Effect on active and passive risk avoidance.
Journal of Behavioral Decision Making, *27*(5).
https://doi.org/10.1002/bdm.1821

Kahneman, D., Knetsch, J. L., & Thaler, R. H.
(2019).
Anomalies: The endowment effect, loss aversion, and status quo bias.
Choices, values, and frames.
https://doi.org/10.1017/CBO9780511803475.009

Kahnt, T., Grueschow, M., Speck, O., & Haynes, J. D.
(2011).
Perceptual Learning and Decision-Making in Human Medial Frontal Cortex.
Neuron, *70*(3).
https://doi.org/10.1016/j.neuron.2011.02.054

Karşılar, H., Simen, P., Papadakis, S., & Balcı, F.
(2014).

Speed Accuracy Trade-off under Response Deadlines.
Procedia - Social and Behavioral Sciences, *126*.
https://doi.org/10.1016/j.sbspro.2014.02.371

Katayama, K., Sakamoto, H., & Narihisa, H.
(2000).
The efficiency of hybrid mutation genetic algorithm for the travelling salesman problem.
Mathematical and Computer Modelling, *31*(10-12),
197–203.
https://doi.org/10.1016/S0895-7177(00)00088-1

Katsimpokis, D., Hawkins, G. E., & van Maanen, L.
(2020).
Not all Speed-Accuracy Trade-Off Manipulations Have the Same Psychological Effect.
Computational Brain & Behavior, *3*(3).
https://doi.org/10.1007/s42113-020-00074-y

Kaufman, H., & Howard, R. A.
(1961).
Dynamic Programming and Markov Processes.
The American Mathematical Monthly, *68*(2).
https://doi.org/10.2307/2312519

Khodadadi, A., Fakhari, P., & Busemeyer, J. R.
(2017).
Learning to allocate limited time to decisions with different expected outcomes.
Cognitive Psychology, *95*.
https://doi.org/10.1016/j.cogpsych.2017.03.002

Kimura, H., Yamamura, M., & Kobayashi, S.
(1995).
Reinforcement Learning by Stochastic Hill Climbing on Discounted Reward.
Machine learning proceedings 1995.
https://doi.org/10.1016/b978-1-55860-377-6.50044-x

Kruschke, J. K.
(2014).
Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan, second edition.
https://doi.org/10.1016/B978-0-12-405888-0.09999-2

Lin, H., Saunders, B., Friese, M., Evans, N. J., & Inzlicht, M.
(2020).
Strong Effort Manipulations Reduce Response Caution: A Preregistered Reinvention of the Ego-Depletion Paradigm.
Psychological Science, *31*(5).

https://doi.org/10.1177/0956797620904990

Litterman, R. B.

(1983).

A random walk, markov model for the distribution of time series.

Journal of Business and Economic Statistics, *1*(2).

https://doi.org/10.1080/07350015.1983.10509336

Malhotra, G., Leslie, D. S., Ludwig, C. J., & Bogacz, R.

(2017).

Overcoming indecision by changing the decision boundary.

Journal of Experimental Psychology: General, *146*(6).

https://doi.org/10.1037/xge0000286

Malhotra, G., Leslie, D. S., Ludwig, C. J., & Bogacz, R.

(2018).

Time-varying decision boundaries: insights from optimality analysis.

Psychonomic Bulletin and Review, *25*(3).

https://doi.org/10.3758/s13423-017-1340-6

Mayrhofer, J. M., Skreb, V., Von der Behrens, W., Musall, S., Weber, B., & Haiss, F.

(2013).

Novel two-alternative forced choice paradigm for bilateral vibrotactile whisker frequency discrimination in head-fixed mice and rats.

Journal of Neurophysiology, *109*(1).

https://doi.org/10.1152/jn.00488.2012

Miletić, S., & van Maanen, L.

(2019).

Caution in decision-making under time pressure is mediated by timing ability.

Cognitive Psychology, *110*.

https://doi.org/10.1016/j.cogpsych.2019.01.002

Moran, R.

(2014).

Optimal decision making in heterogeneous and biased environments.

Psychonomic Bulletin and Review, *22*(1).

https://doi.org/10.3758/s13423-014-0669-3

Myung, I. J., & Busemeyer, J. R.

(1989).

Criterion Learning in a Deferred Decision-Making Task.

The American Journal of Psychology, *102*(1).

https://doi.org/10.2307/1423113

Oud, B., Krajbich, I., Miller, K., Cheong, J. H., Botvinick, M., & Fehr, E.

(2016).
Irrational time allocation in decision-making.
Proceedings of the Royal Society B: Biological Sciences, *283*(1822).
https://doi.org/10.1098/rspb.2015.1439

Padel, S., & Foster, C.
(2005).
Exploring the gap between attitudes and behaviour: Understanding why consumers buy or do not buy organic food.
British Food Journal, *107*(8).
https://doi.org/10.1108/00070700510611002

Palestro, J. J., Weichart, E., Sederberg, P. B., & Turner, B. M.
(2018).
Some task demands induce collapsing bounds: Evidence from a behavioral analysis.
Psychonomic Bulletin and Review, *25*(4).
https://doi.org/10.3758/s13423-018-1479-9

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K.
(2019).
PsychoPy2: Experiments in behavior made easy.
Behavior Research Methods, *51*(1).
https://doi.org/10.3758/s13428-018-01193-y

Pitz, G. F., & Reinhold, H.
(1968).
Payoff effects in sequential decision-making.
Journal of Experimental Psychology, *77*(2).
https://doi.org/10.1037/h0025802

R Core Team.
(2019).
R: A language and environment for statistical computing.

Ratcliff, R.
(1978).
A theory of memory retrieval.
Psychological Review, *85*(2).
https://doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R., Cherian, A., & Segraves, M.
(2003).
A Comparison of Macaque Behavior and Superior Colliculus Neuronal Activity to Predictions From Models of Two-Choice Decisions.

Journal of Neurophysiology, *90*(3).
https://doi.org/10.1152/jn.01049.2002

Reckless, G. E., Bolstad, I., Nakstad, P. H., Andreassen, O. A., & Jensen, J.
(2013).
Motivation alters response bias and neural activation patterns in a perceptual decision-making task.
Neuroscience, *238*.
https://doi.org/10.1016/j.neuroscience.2013.02.015

Schubert, A. L., Frischkorn, G. T., Hagemann, D., & Voss, A.
(2016).
Trait characteristics of diffusion model parameters.
Journal of Intelligence, *4*(3).
https://doi.org/10.3390/jintelligence4030007

Schultz, W.
(2006).
Behavioral theories and the neurophysiology of reward.
Annual Review of Psychology, *57*.
https://doi.org/10.1146/annurev.psych.56.091103.070229

Senftleben, U., Schoemann, M., Rudolf, M., & Scherbaum, S.
(2021).
To stay or not to stay: The stability of choice perseveration in value-based decision making.
Quarterly Journal of Experimental Psychology, *74*(1).
https://doi.org/10.1177/1747021820964330

Senftleben, U., Schoemann, M., Schwenke, D., Richter, S., Dshemuchadse, M., & Scherbaum, S.
(2019).
Choice perseveration in value-based decision making: The impact of inter-trial interval and mood.
Acta Psychologica, *198*.
https://doi.org/10.1016/j.actpsy.2019.102876

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E. J.
(2008).
A survey of model evaluation approaches with a tutorial on hierarchical bayesian methods.
Cognitive Science, *32*(8).
https://doi.org/10.1080/03640210802414826

Simen, P., Contreras, D., Buck, C., Hu, P., Holmes, P., & Cohen, J. D.
(2009).

Reward Rate Optimization in Two-Alternative Decision Making: Empirical Tests of Theoretical Predictions.
Journal of Experimental Psychology: Human Perception and Performance, *35*(6).
https://doi.org/10.1037/a0016926

Simon, H. A.
(1972).
Theories of bounded rationality.
Decision and organization, *1*.

Simon, H. A.
(1955).
A behavioral model of rational choice.
Quarterly Journal of Economics, *69*(1).
https://doi.org/10.2307/1884852

Smith, P. L., & Ratcliff, R.
(2004).
Psychology and neurobiology of simple decisions.
https://doi.org/10.1016/j.tins.2004.01.006

Starns, J. J., & Ratcliff, R.
(2010).
The effects of aging on the speed-accuracy compromise: Boundary optimality in the diffusion model.
Psychology and Aging, *25*(2).
https://doi.org/10.1037/a0018022

Stich, K. P., & Winter, Y.
(2006).
Lack of generalization of object discrimination between spatial contexts by a bat.
Journal of Experimental Biology, *209*(23).
https://doi.org/10.1242/jeb.02574

Summerfield, C., Behrens, T. E., & Koechlin, E.
(2011).
Perceptual classification in a rapidly changing environment.
Neuron, *71*(4).
https://doi.org/10.1016/j.neuron.2011.06.022

Sutton, R., & Barto, A.
(1998).
Reinforcement Learning: An Introduction.
IEEE Transactions on Neural Networks, *9*(5).
https://doi.org/10.1109/tnn.1998.712192

Thompson, W. R.
(1933).
On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples.
Biometrika, *25*(3/4).
https://doi.org/10.2307/2332286

Trueblood, J. S., Heathcote, A., Evans, N. J., & Holmes, W. R.
(2021).
Urgency, leakage, and the relative nature of information processing in decision-making.
Psychological Review, *128*(1).
https://doi.org/10.1037/rev0000255

Tversky, A., & Kahneman, D.
(1974).
Judgment under uncertainty: Heuristics and biases.
Science, *185*(4157).
https://doi.org/10.1126/science.185.4157.1124

Vehtari, A., Gelman, A., & Gabry, J.
(2017).
Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC.
Statistics and Computing, *27*(5).
https://doi.org/10.1007/s11222-016-9696-4

Voskuilen, C., Ratcliff, R., & Smith, P. L.
(2016).
Comparing fixed and collapsing boundary versions of the diffusion model.
Journal of Mathematical Psychology, *73*.
https://doi.org/10.1016/j.jmp.2016.04.008

White, C. N., Mumford, J. A., & Poldrack, R. A.
(2012).
Perceptual criteria in the human brain.
Journal of Neuroscience, *32*(47).
https://doi.org/10.1523/JNEUROSCI.1744-12.2012

Wischnewski, M., & Schutter, D. J.
(2018).
Dissociating absolute and relative reward- and punishment-related electrocortical processing: An event-related potential study.
International Journal of Psychophysiology, *126*,
13–19.
https://doi.org/10.1016/j.ijpsycho.2018.02.010

Xu, S., Pan, Y., Qu, Z., Fang, Z., Yang, Z., Yang, F., Wang, F., & Rao, H.
(2018).
Differential effects of real versus hypothetical monetary reward magnitude on risk-taking behavior and brain activity.
Scientific Reports, *8*(1).
https://doi.org/10.1038/s41598-018-21820-0

Yu, A. J., & Cohen, J. D.
(2009).
Sequential effects: Superstition or rational behavior?
Advances in Neural Information Processing Systems 21 - Proceedings of the 2008 Conference.

Zacksenhouse, M., Bogacz, R., & Holmes, P.
(2010).
Robust versus optimal strategies for two-alternative forced choice tasks.
Journal of Mathematical Psychology, *54*(2).
https://doi.org/10.1016/j.jmp.2009.12.004

Zhang, R., Brennan, T. J., & Lo, A. W.
(2014).
The origin of risk aversion.
Proceedings of the National Academy of Sciences of the United States of America, *111*(50).
https://doi.org/10.1073/pnas.1406755111