Situation-appropriate Investment of Cognitive Resources

DISSERTATION

for the degree of

Doctor rerum naturalium
(Dr. rer. nat.)

submitted to

the School of Science
at Technische Universität Dresden

by

Florian Ott

born on 15.07.1989 in Rottweil

defended on 21.03.2022

This dissertation thesis was completed from November 2016 to December 2021 at the Institute of General Psychology, Biopsychology and Methods of Psychology at the chair of Neuroimaging under the supervision of Prof. Dr. Stefan Kiebel and Prof. Dr. Alexander Strobel

Acknowledgments

First and foremost, I would like to thank my primary supervisor Stefan Kiebel for his continuous support and encouragement during my doctoral studies and for the inspiring experience of being a part of the Chair of Neuroimaging. I am also grateful to my secondary supervisor Alexander Strobel for his support and for coordinating the integrated research training group of the CRC 940, which always provided a good platform to meet other doctoral students and learn new skills. My thanks extend to my colleagues Dimitrije Marković and Sebastian Bitzer for their valuable support and advice on all the tricky modelling issues and to Cassandra Visconti, Dario Cuevas Rivera, Eric Legler, Sascha Frölich and Sarah Schwöbel for the numerous scientific discussions that were not only insightful but also fun! I am deeply grateful to my parents for their constant loving support all through my studies. A very special thank you goes to my partner Elisabeth without whom it would have been impossible for me to complete this dissertation.

List of publications used in this thesis

Ott, F., Marković, D., Strobel, A., & Kiebel, S. J. (2020). Dynamic integration of forward planning and heuristic preferences during multiple goal pursuit. *PLoS computational biology*, *16*(2), e1007685.

Chapter 2 is based on this publication.

Author contributions according to the Contributor Role Taxonomy (CRediT):
Florian Ott: Conceptualisation, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Visualisation, Writing – Original Draft, Writing – Review and Editing. *Dimitrije Marković:* Methodology, Software, Validation, Writing – Original Draft, Writing – Review and Editing. *Alexander Strobel:* Conceptualisation, Supervision, Funding Acquisition, Writing – Review and Editing. *Stefan Kiebel:* Conceptualisation, Supervision, Funding Acquisition, Writing – Original Draft, Writing – Review and Editing.

Data availability: To allow for maximum transparency and reproducibility, the datasets and analysis code associated with this publication are publicly available at https://doi.org/10.5281/zenodo.5526426

Exclusive use: This publication is not currently used in any other dissertation, nor is it intended to be used in any future dissertations.

Ott, F., Legler, E., & Kiebel, S. J. (2021). Forward planning driven by context dependent conflict processing in anterior cingulate cortex. bioRxiv 2021. https://biorxiv.org/content/10.1101/2021.07.19.452905

Chapter 3 is based on this preprint.

Author contributions according to the Contributor Role Taxonomy (CRediT):
Florian Ott: Conceptualisation, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Visualisation, Writing – Original Draft. *Eric Legler*: Investigation, Writing – Review and Editing. *Stefan Kiebel*: Conceptualisation, Supervision, Funding Acquisition, Writing – Review and Editing

Data availability: To allow for maximum transparency and reproducibility, the datasets and analysis code associated with this publication are publicly available at https://doi.org/10.5281/zenodo.5112965

Exclusive use: This publication is not currently used in any other dissertation, nor is it intended to be used in any future dissertations.

Contents

List of Figures

List of Tables

Abstract

The human brain is equipped with the ability to plan ahead, i.e. to mentally simulate the expected consequences of candidate actions to select the one with the most desirable expected long-term outcome. Insufficient planning can lead to maladaptive behaviour and may even be a contributory cause of important societal problems such as the depletion of natural resources or man-made climate change. Understanding the cognitive and neural mechanisms of forward planning and its regulation are therefore of great importance and could ultimately give us clues on how to better align our behaviour with long-term goals.

Apart from its potential beneficial effects, planning is time-consuming and therefore associated with opportunity costs. It is assumed that the brain regulates the investment into planning based on a cost-benefit analysis, so that planning only takes place when the perceived benefits outweigh the costs. But how can the brain know in advance how beneficial or costly planning will be? One potential solution is that people learn from experience how valuable planning would be in a given situation. It is however largely unknown how the brain implements such learning, especially in environments with large state spaces.

This dissertation tested the hypothesis that humans construct and use so-called control contexts to efficiently adjust the degree of planning to the demands of the current situation. Control contexts can be seen as abstract state representations, that conveniently cluster together situations with a similar demand for planning. Inferring context thus allows to prospectively adjust the control system to the learned demands of the global context. To test the control context hypothesis, two complex sequential decision making tasks were developed. Each of the two tasks had to fulfil two important criteria. First, the tasks should generate both situations in which planning had the potential to improve performance, as well as situations in which a simple strategy was sufficient. Second, the tasks had to feature rich state spaces requiring participants to compress their state representation for efficient regulation of planning. Participants' planning was modelled using a parametrized dynamic programming solution to a Markov Decision Process, with parameters estimated via hierarchical Bayesian inference.

The first study used a 15-step task in which participants had to make a series of decisions to achieve one or multiple goals. In this task, the computational costs of accurate forward planning increased exponentially with the length of the planning horizon. We therefore hypothesized that participants identify 'distance from goal' as the relevant contextual feature to guide their regulation of forward planning. As expected we found that participants predominantly relied on a simple heuristic when still far from the goal but progressively switched towards forward planning when the goal approached.

In the second study participants had to sustainably invest a limited but replenishable energy resource, that was needed to accept offers, in order to accumulate a maximum number of points in the long run. The demand for planning varied across the different situations of the task, but due to the large number of possible situations (n = 448) it would be difficult for the participants to develop an expectation for each individual situation of how beneficial planning would be. We therefore hypothesized, that to regulate their forward planning participants used a compressed tasks representation, clustering together states with similar demands for planning. Consistent with this, reaction times (operationalising planning duration) increased with trial-by-trial value-conflict (operationalising approximate planning demand), but this increase was more pronounced in a context with generally high demand for planning. We further found that fMRI activity in the dorsal anterior cingulate cortex (dACC) increased with conflict, but this increase was more pronounced in a context with generally high demand for planning as well. Taken together, the results suggest that the dACC integrates representations of planning demand on different levels of abstraction to regulate prospective information sampling in an efficient and situation-appropriate way.

This dissertation provides novel insights into the question how humans adapt their planning to the demands of the current situation. The results are consistent with the view that the regulation of planning is based on an integrated signal of the expected costs and benefits of planning. Furthermore, the results of this dissertation provide evidence that the regulation of planning in environments with real-world complexity critically relies on the brain’s powerful ability to construct and use abstract hierarchical representations.

1 General introduction

1.1 Cognitive control - Benefits, costs and the necessity of regulation

In a familiar and stable context, we often have well-proven default behaviours at our disposal that we can carry out without much deliberation. However, changes in our motivation or in external conditions may render these default behaviours inappropriate. Cognitive control allows us to disengage from these default behaviours, process additional information and flexibly reconfigure our behaviour to meet the demands of the situation. (Goschke, 2013; E. K. Miller & Cohen, 2001). Consider the example of a ride to work by bicycle. On a regular day the rider just follows the default route, navigating through the streets almost automatically. However, should the road be blocked due to an accident, the rider has to find a new route, requiring to process additional information including the mental simulation of potential alternatives.

Clearly, cognitive control has the potential to improve performance in many situations, especially in situations that are novel or uncertain and if current action has great consequences for future possibilities. So why do we not constantly employ the greatest possible degree of cognitive control? The reason why cognitive control has to be regulated is that it incurs an opportunity cost (Kurzban et al., 2013; Shenhav et al., 2017). For humans with their limited processing capacities, engaging cognitive control takes some computation time, as it entails processes like the gating of contextual information into working memory, the anticipation of future action consequences or the temporal integration of different information sources. The opportunity costs thus result from the behaviours and cognitive activities that are foregone due to the employment of the control process. A simple example would be when one arrives at a restaurant just before closing time and begins to carefully examine the menu to make the selection that best suits one's current desires. In this scenario the kitchen would already be closed after the main course and there would be no time to order dessert. Another option would have been to choose a familiar main course quickly without much consideration, so there would be enough time to order a subsequent dessert. In this example, the advantages of deliberating about the main course are offset by the risk of having to forego a dessert.

Recent work suggests that cognitive control should be allocated to optimize a cost-benefit trade-off (Boureau et al., 2015; Gershman et al., 2015; Griffiths et al., 2019; Keramati et al., 2011; Shenhav et al., 2013). It seems that this would require to estimate the performance improvement, but also the costs of a controlled strategy compared to the standard strategy. Paradoxically, however, the exact calculation of the benefits and costs of control would incur its own computational costs. It is therefore an obvious thought that the brain uses other mechanisms to circumvent this elaborate or even impossible calculation. One possibility would be the use of readily available and approximate signals to decide about when and what kind of control to engage.

1.2 Learning cognitive control

One way of avoiding the explicit calculation of control demands is to conceive the regulation of control (i.e. meta-control) as a process of learning (Abrahamse et al., 2016; Chiu & Egner, 2019; Egner, 2014; Lieder et al., 2018). According to this perspective, a particular state of the control system can become bound to a stimulus or context. Later, after learning, contextual cues can rapidly trigger the control processes that have been associated with that context in the past. This learning is sensitive to reward information and potentially includes a mapping from the context to a representation of the value of control (Lieder et al., 2018). Based on such a learned mapping, the brain could thus select an appropriate control process for a given context without the need for an explicit calculation of costs and benefits. An accumulating body of research is consistent with a value-based learning account of cognitive control, providing evidence for cross-context transfer of learned control strategies and for context-dependent effects of reward and costs on a range of different control functions (for recent reviews see M. Botvinick & Braver, 2015; Eppinger et al., 2021).

1.2.1 Introducing basic findings in cognitive control tasks

Modulatory effects of reward on conflict processing have been observed in classic response interference tasks, such as the Stroop, Flanker or Simon task. These tasks simultaneously evoke two stimulus-response mappings of which one is typically prepotent and more automatic. In some of the experiments' trials, the two mappings are congruent, whereas in others, they are incongruent and prescribe conflicting responses. In such conflict trials, successful task performance requires to supress the prepotent response and to activate the task-relevant stimulus-response mapping. This process of reconfiguration and associated attentional processes are thought to explain the typical finding that people take more time to respond and are more likely to commit an error during conflict trials compared to congruent trials (Eriksen & Eriksen, 1974; J. R. Simon, 1969; Stroop, 1935). The adjustment of the brain networks towards a more controlled processing mode in incongruent trials has been shown to improve conflict resolution in following incongruent trials. Behaviourally, this congruency sequence effect is reflected in that the slowing of responses and the increase in error rate in incongruent trials is reduced, if the previous trial was also incongruent (Gratton et al., 1992).

1.2.2 Control is modulated by reward context

Importantly, previous studies showed that conflict processing can be modulated by reward incentives and contextual manipulations of control demand. In a modified version of the Stroop task where colour naming was associated with a potential reward for a subset of ink colours, congruency effects were reduced in potentially rewarded trials relative to unrewarded trials (Krebs et al., 2013; Krebs et al., 2010). The authors concluded that anticipation of reward promotes effective stimulus processing by increasing attention to the reward predictive stimulus, thus resulting in lower error rates and faster responding, i.e. a decreased congruency effect. Padmala and Pessoa (2011) used a slightly different variant of the Stroop task in which participants were informed of the possible reward by a cue phase prior to the actual Stroop task. They as well found a decreased congruency effect in rewarded trials, suggesting that the prospect of reward can proactively prime the control system to invest cognitive resources in conflict resolution. Recently, it was proposed that reward effects on cognitive control in incentivized Stroop tasks can be explained by a mechanism that learns to predict the value of control based on a linear combination of features in a situation (Bustamante et al., 2021). Interestingly, these authors' model also explains how learned control demands generalize to novel situations based on a shared set of features and of how this transfer may lead to a suboptimal allocation of control if nonlinear combinations of features are predictive of control demand. The aforementioned congruence sequence effect is another important behavioural marker for cognitive control that can be modulated by reward. Previous studies showed that the performance improvement in an incongruent trial which was preceded by another incongruent trial can be enhanced if the preceding trial was rewarded (Braem et al., 2012; Stürmer, 2011). Additionally, Stürmer (2011) found that the typical response slowing after errors was increased in a context where reward could be obtained.

1.2.3 Control is modulated by difficulty context

How much cognitive resources are invested depends not only on the expected reward but also on the expected difficulty in a particular context. Previous studies modulated context difficulty in a Stroop task by combining one part of the task-relevant stimuli with mostly incongruent task-irrelevant features and the other part with mostly congruent task-irrelevant features (Bugg & Dey, 2018; Bugg et al., 2011). The results of these studies showed a proportion congruency effect characterized by slower reaction times for incongruent trials in a mostly congruent context compared to a mostly incongruent context. It was further shown that adjustments in control putatively driven by learned stimulus-congruency associations transferred to novel exemplars of the same stimulus category. Another study extended these findings, providing evidence for the transfer of stimulus-control associations across arbitrarily linked stimuli (Bejjani et al., 2018). In this study the authors used a Stroop task with item specific congruency frequency manipulations and three distinct experimental phases. First, in the stimulus-stimulus association phase, specific face or house images were preceded by a particular scene stimulus, such that pairs of two were formed. Second, in the stimulus-control learning phase, particular scene images were followed by mostly congruent or mostly incongruent Stroop trials. And third, in the stimulus control transfer phase, face and house images were followed by Stroop trials. As expected, they found that the congruency expectations learned for scenes, transferred to their associated face and house images as evidenced by a proportion congruency effect.

The degree to which people invest in cognitive control depends on expectations of how beneficial control will be in a given context. Taken together, the reviewed empirical evidence supports the notion that the allocation of control is regulated at least partially by learned associations between a structured representational stimulus space and an integrated expected value of control.

1.3 Regulation of forward planning

1.3.1 Conceptualizing planning as information sampling

In conflict situations, one is usually uncertain what to do. It is therefore often necessary to sample additional information to gain certainty about which action best subserves one's long-term goals. Additional information can be sampled both overtly and covertly (L. Hunt et al., 2021). Overt information sampling can be realized by either actively directing one's gaze to relevant sources of information or by exploratory actions aimed specifically at gaining knowledge about the environment. In the case of covert information sampling, however, information is generated internally by activating representations from memory. A special form of covert information sampling is the mental simulation of future states, actions and their consequences (Schacter et al., 2012). This process is commonly referred to as forward planning and is a central component of cognitive control (M. M. Botvinick & Cohen, 2014; E. K. Miller & Cohen, 2001), along with the basic inhibitory and attentional processes examined in classical interference tasks. Planning ahead can be particularly beneficial when the actions we take in the present have a strong influence on future states and action opportunities. Examples illustrating the benefits of anticipating future consequences are diverse and range from health-promoting behaviour to the sustainable use of limited natural resources, to the coordination of a sequence of interdependent actions during spatial navigation. Besides the beneficial effects of planning, there is also a downside, namely that planning takes time and is thus associated with opportunity costs.

As motivated above, it is thus crucial for the brain to consider both the costs and benefits for control allocation. This becomes especially evident when deciding about how long to plan ahead. However, finding the optimal cost-benefit trade-off is hard, since it is often not known in advance how resources invested in planning will improve subsequent decisions. The question of how the brain balances costs and benefits of planning is an active area of research (e.g. Piray & Daw, 2021), but one can identify at least two different forms of regulation acting in parallel. First, the regulation of planning could be conceived of as a emergent property within a dynamic process of information sampling (Krajbich et al., 2010). And second, a superordinate process might prospectively adjust the parameters of this sampling process to the expected demands of the current situation (D. G. Lee & Daunizeau, 2021; Lieder et al., 2018). In the following I will briefly review evidence for both reactive and prospective control of planning to lay the basis for formulating the hypotheses of this dissertation.

1.3.2 The regulation of planning emerges from a dynamic process of information sampling

A standard model for information sampling during decision making is the drift diffusion model (DDM, Ratcliff et al., 2016). Traditionally, the DDM has been applied to perceptual decision making (Gold & Shadlen, 2007), but more recently it has also been applied value-based decision making, i.e. when one has to judge which of several options one prefers (Busemeyer et al., 2019; Tajima et al., 2016). In the DDM, a response is generated if accumulated samples of noisy evidence reach a predefined threshold. The DDM models the dynamics of the decision process and thus allows for both predictions of choice outcome and response times. In its simplest form a DDM can be characterized by five parameters: The threshold specifies the amount of evidence that must be accumulated until a response is generated. Higher thresholds can lead to more accurate decisions but at cost of longer response times. The drift rate specifies the speed with which the threshold is approached. Importantly, in value-based DDMs with two alternatives the drift rate is usually a function of the value difference (Krajbich et al., 2010; Mormann et al., 2010). If the value difference is small, decision difficulty (or conflict) is considered high, resulting in slower drift rates and thus slower response times. Further parameters specify the noise in the diffusion process, the non-decision time (e.g. for execution of the motor signal) and an initial bias for one of the options. How much information is sampled could thus be controlled dynamically in the following way: First, rough prior estimates of option values and uncertainties are assessed. If there is no clear evidence for one or the other option (i.e. if the value difference is small and value estimates uncertain), additional information is sampled, which consequentially modulates value estimates and their uncertainties. This process continues until a decisive amount of evidence for one over the other option is found. Empirical research showed that diffusion models can account for choice, response times, and neural activity during simple value-based decision making (Blair et al., 2006; De Martino et al., 2013; L. T. Hunt et al., 2012; Krajbich et al., 2010; Mormann et al., 2010; Pochon et al., 2008).

In contrast to single-trial economic decision making tasks, behaviour in the real world is goal-directed and temporally extended. Making good decisions therefore often affords to not only evaluate an action's immediate value but also its impact on the value of potential future actions. In other words, goal-directed behaviour often requires planning ahead. In such cases the sampling of evidence, as e.g. modelled with the DDM, involves the memory-dependent mental simulation of action sequences and their expected outcomes (Biderman et al., 2020; Shadlen & Shohamy, 2016; Wang et al., 2020). Consistent with this account, response times correlate with choice conflict in sequential decision making tasks which do afford forward planning (Korn & Bach, 2018; Shenhav et al., 2014). Further useful characterisation of the planning process has been given by model-based reinforcement learning (Collins & Cockburn, 2020; Dolan & Dayan, 2013; Sutton & Barto, 2018). In model-based reinforcement learning humans are assumed to acquire a model of their environment, including state transition and reward functions, which they can then use to flexibly sample sequences of actions and their outcomes (analogous to a search through the decision tree). Using a sequential two-step task, Doll et al. (2015) provided linking evidence that such model-based decisions are based on memory-dependent prospective sampling, involving the hippocampus.

1.3.3 Prospective and value-based regulation of planning

Besides the reviewed evidence indicating that the regulation of planning emerges from a dynamic process of information sampling, there is also evidence for a more prospective mode of regulation (D. G. Lee & Daunizeau, 2021; Lieder et al., 2018). For basic cognitive control functions we already discussed in section 1.2 that people prepare their cognitive systems for the upcoming requirements based on learned context-control associations. It has been suggested that these learning processes play also an important role for the regulation of more complex control functions such as planning. For example, Lieder et al. (2018) suggested that people could adequately adjust the height of a threshold for information sampling based on a learned mapping between a context and an integrated value (including costs and benefits) associated with the threshold setting. Such context-based learning has least two advantages. First, based on previously learned context-control associations, the control system could be proactively adjusted to the planning demands of a novel situation in which it is not clear initially how efficiently information can be obtained. Second, situations in which planning will likely not improve performance can be excluded a priori from further evaluation, thus reducing overall cost of planning. Empirical evidence supports this control learning perspective, showing that people can learn to deliberate more in a context where it pays off (Lieder & Griffiths, 2017). However, more studies are needed to directly test the role of context-control learning during planning.

Findings from the model-based reinforcement learning literature are also consistent with a context- and value-dependent learning account of planning regulation. In a multistep-decision task, Kool et al. (2017) showed that the prospect of reward increases the propensity to plan ahead, but only if planning promises higher reward relative to a simpler model-free strategy. Complementary to that, another study showed that people become less likely to invest in effortful planning if planning is associated with a greater cost, operationalized as increased depth of the causal tasks structure (Kool et al., 2018). Further evidence suggests that people can gradually adapt the speed-accuracy trade-off of planning to the current situation, for example by limiting the planning horizon (Juechems et al., 2019; Keramati et al., 2016), i.e. the number of future steps considered, or by selectively evaluating only part of the decision tree (Huys et al., 2012).

1.4 The role of the anterior cingulate cortex in deciding about the investment of cognitive resources

The dACC is part of a network that enables adaptive and flexible responding in cognitively demanding situations, i.e. situations that require cognitive control (Duncan, 2010; Niendam et al., 2012). The exact function that the dACC plays in these situations is not fully understood and a matter of ongoing research (Ebitz & Hayden, 2016; Heilbronner & Hayden, 2016). One classic account suggests that the dACC plays a central role in the regulation of cognitive control by monitoring processing conflicts that serve as a signal for control demand (M. M. Botvinick et al., 2001). Whereas, according to this account, a monitoring function is ascribed to the dACC, the actual implementation of control is thought to be carried out by a network of other cortical and subcortical structures. Empirical research supports the role of the dACC in conflict monitoring, showing the dACC encodes response conflict in interference tasks like the Stroop, Simon or Flanker tasks (Kerns et al., 2004; MacDonald et al., 2000; E. H. Smith et al., 2019). In line with these findings, recent electrophysiological studies on monkeys imply that the monkey dACC may also contain explicit representations of uncertainty that support the control of information sampling during reward-based decision making (Monosov et al., 2020; White et al., 2019). However, previous research indicated that the dACC does not only track conflict and uncertainty but also signals reflecting the short- and long-term value of choice options (Heilbronner & Hayden, 2016; Kolling et al., 2018) .

During value-based decision making, conflict arises when there is no clear preference for one option over another option. To resolve this conflict it is often necessary to sample additional information either internally from memory or externally from the environment. It is widely assumed that the dACC acts as a central bottleneck that controls such additional information sampling in the face of conflict. Mechanistically, this could be implemented by inhibiting prepotent response tendencies via a hyperdirect pathway from the dACC to the subthalamic nucleus (STN) of the basal ganglia (Frank, 2006; Frank et al., 2015; Jahfari et al., 2011; Wiecki & Frank, 2013; Wiecki et al., 2013). This would buy more time (i.e. raising a decision threshold) for controlled processing, like prospective value-based information sampling, to influence response generation. Empirical evidence supports a conflict monitoring role of the dACC during value-based decision making, showing dACC activity correlates with the absolute difference between options values (i.e. conflict) during value-based decision making (Blair et al., 2006; Hare et al., 2011; Pochon et al., 2008; Venkatraman et al., 2009). Further studies suggest more specifically, that activity in the dACC controls additional information sampling via the conflict-dependent modulation of a decision threshold (Cavanagh et al., 2011; Frank et al., 2015; Gluth et al., 2012). The dACC is also involved in controlling more sophisticated forms of information sampling like planning ahead. In sequential decision making tasks that afford such planning, dACC activity has been found to increase with measures of conflict and choice uncertainty (Economides et al., 2015; Kolling et al., 2014; Korn & Bach, 2018; Schwartenbeck et al., 2015; Shenhav et al., 2014).

1.5 Hypotheses

Cognitive control and planning in particular is costly, and therefore must be regulated, such that the amount of cognitive resources invested is adequate to the current situation (see section 1.1). However, knowing in advance how beneficial forward planning will be in a given situation is hard. The only way to know the exact value of planning would be to actually do it, which would ab initio defeat the purpose of regulating planning, i.e. the reduction of computational and time costs. One possible solution to this dilemma is that the allocation of control is regulated by learned associations between stimuli and control network configurations (see sections 1.2 and 1.3.3). Such learning likely includes generalisation processes that cluster together stimulus states with similar control relevant properties into more general control contexts. With that, the brain could infer the demand for control, based on previous experience with situations that share some structural properties with the current situation.

This dissertation addressed the question of how people use control contexts to efficiently balance the benefits and costs of investing cognitive resources. The focus was specifically on how people invest resources in planning multiple steps into the future, because the high computational costs involved in planning particularly motivate its regulation. This question was addressed using two newly developed complex sequential decision making tasks along with cognitive computational modelling and model-based fMRI. Each of the two developed tasks had to fulfil two key requirements. First, the tasks had to include both situations in which forward planning could improve performance and situations in which a simple heuristic was sufficient. Second, the tasks had to feature large state spaces providing participants with the requirement to compress their tasks representation in order to decide efficiently about their planning.

The first behavioural study tested how the mixing of forward planning with simple heuristics changes when people progress in a goal-reaching scenario with a fixed deadline. We used a sequential task in which participants accumulated two different types of points (A- and B-points) by accepting or rejecting offers. If points surpassed a point-specific threshold after the fixed deadline, the A- or B goal counted as achieved and a monetary reward was given. Consequentially, there were four different goal outcomes: None of the goals was achieved, either goal A or goal B was achieved or both goals were achieved, with the latter result yielding twice the reward than if only one goal was achieved. Participants thus had to decide on a trial-by-trial basis whether they should focus on only one goal or whether they should try to promote both goals simultaneously. Deciding optimally which strategy to pursue, given the current amount of A- and B-points and the number of remaining action opportunities, would require to plan ahead multiple time steps (trials) until the end of the deadline. However, in this study, the computational costs associated with such planning explodes exponentially with the length of the planning horizon. We therefore hypothesized that participants would use a simple heuristic when the deadline is temporally distant and planning costs are prohibitively high, and progressively shift towards forward planning when the deadline approaches and computational costs become affordable. We specifically assumed that this shift would be driven by participants identifying and using “distance from deadline” as the relevant contextual indicator for the value to plan ahead.

The second fMRI study investigated how inferred control contexts facilitate the situation-appropriate investment into forward planning via a contextually modulated processing of trial-by-trial conflicts in the dACC. To address this question, a complex sequential decision making task was developed, in which participants had to sustainably invest a limited but replenishable energy resource, that was needed to accept offers, in order to accumulate a maximum number of points in the long run. Clearly, neither a greedy strategy, i.e. to invest all the energy immediately, nor an overly conservative strategy, i.e. to save all the energy for high offers only would lead to an optimal outcome. Deciding optimally would thus require to plan ahead multiple steps to anticipate the consequences of energy consumption on future action opportunities. Importantly, however, the utility of such planning varied across the different situations encountered in the task. And because planning is typically perceived as costly, we expected that participants adapt the degree of planning to its changing utility. Knowing in advance for every situation how beneficial planning will be was difficult, because of the complexity of the task (n_{States} = 448). We therefore hypothesized that, to determine their cognitive resource investment, participants leveraged a generalized task space (i.e. control contexts) that grouped together states with a similar demand for planning. Control contexts could then be used to prospectively reduce, for a subset of situations, the degree of extended evaluation by planning. We specifically hypothesized that response times (as a behavioural marker for planning) increase with conflict (operationalised as the difference between action values), but that this increase is more pronounced, if the participants inferred to be in a context of high control demand. We further tested whether the context dependency of the coupling between conflict and planning could be mediated by context-dependent conflict processing in the ACC. We predicted that activity in the ACC increases with conflict but, that this increase is more pronounced in a context with generally high planning demand.

2 Study 1: Dynamic integration of forward planning and heuristic preferences during multiple goal pursuit

2.1 Abstract

Selecting goals and successfully pursuing them in an uncertain and dynamic environment is an important aspect of human behaviour. In order to decide which goal to pursue at what point in time, one has to evaluate the consequences of one's actions over future time steps by forward planning. However, when the goal is still temporally distant, detailed forward planning can be prohibitively costly. One way to select actions at minimal computational costs is to use heuristics. It is an open question how humans mix heuristics with forward planning to balance computational costs with goal reaching performance. To test a hypothesis about dynamic mixing of heuristics with forward planning, we used a novel stochastic sequential two-goal task. Comparing participants' decisions with an optimal full planning agent, we found that at the early stages of goal-reaching sequences, in which both goals are temporally distant and planning complexity is high, on average 42% (SD = 19%) of participants' choices deviated from the agent's optimal choices. Only towards the end of the sequence, participant's behaviour converged to near optimal performance. Subsequent model-based analyses showed that participants used heuristic preferences when the goal was temporally distant and switched to forward planning when the goal was close.

2.2 Author summary

When we pursue our goals, there is often a moment when we recognize that we did not make the progress that we hoped for. What should we do now? Persevere to achieve the original goal, or switch to another goal? Two features of real-world goal pursuit make these decisions particularly complex. First, goals can lie far into an unpredictable future and second, there are many potential goals to pursue. When potential goals are temporally distant, human decision makers cannot use an exhaustive planning strategy, rendering simpler rules of thumb more appropriate. An important question is how humans adjust the rule of thumb approach once they get closer to the goal. We addressed this question using a novel sequential two-goal task and analysed the choice data using a computational model which arbitrates between a rule of thumb and accurate planning. We found that participants' decision making progressively improved as the goal came closer and that this improvement was most likely caused by participants starting to plan ahead.

2.3 Introduction

Decisions of which goal to pursue at what point in time are central to everyday life (Neal et al., 2017; A. M. Schmidt & DeShon, 2007; A. M. Schmidt & Dolis, 2009). Typically, in our dynamic environment, the outcomes of our decisions are stochastic and one cannot predict with certainty whether a preferred goal can be reached. Often, our environment also presents alternative goals that may be less preferred but can be reached with a higher probability than the preferred goal. For example, when working towards a specific dream position in a career, it may turn out after some time that the position is unlikely to be obtained, while another less preferred position can be secured. The decision to make is whether one should continue working towards the preferred position, or switch goals and secure the less preferred position. The risk when pursuing the preferred position is to lose out on both positions. This decision dilemma 'should I risk it and go after a big reward or play it safe and gain less?' is typical for many decisions we have to make in real life. Critically, for many such decisions, these binary choices do not emerge suddenly and unexpectedly, but the decision maker is typically confronted with such decisions after some prolonged period of time working towards enabling different options.

How would one choose one's actions during such a prolonged goal-reaching decision making sequence? One way, if the rules of the dynamic environment and its uncertainties are known, is to use forward planning to always choose the actions which maximize the gain (see Hayes-Roth & Hayes-Roth, 1979; Schacter et al., 2012 reviewing cognitive processes of forward planning). This would be the way one would program an optimal agent in a game or experimental task environment. This approach is often used in cognitive neuroscience to model the mechanism of how humans make decisions in temporally extended goal-reaching scenarios, (e.g. Ballard et al., 2016; Economides et al., 2014; Kolling et al., 2014; Schwartenbeck et al., 2015).

However, the implicit assumption made in these decision making models, namely that humans use detailed forward planning and compute the probabilities of reaching the goals, is difficult to justify, because of the involved computational complexity. In a stochastic environment, forward planning in artificial agents is typically achieved via sampling many possible policies (sequences of actions) which requires substantial computing power that scales exponentially with the number of future actions. In particular, when one is still temporally far from the goal, the computational burden of simulating trajectories into the future is the largest, while the usefulness of the resulting action selection is minimal: intuitively, in stochastic and sufficiently complex environments, anything may yet happen on the long way to the goal so the gain of planning ahead at high cost may be small. The importance of the balance between the benefits and its costs to better understand human decision making became a recent research focus, (e.g. Boureau et al., 2015; Gershman et al., 2015; Lieder & Griffiths, 2017; Shenhav et al., 2013; Shenhav et al., 2017). The question is how one can select actions over long stretches of time, without being exposed to the computational burden of forward planning or similar dynamic programming schemes.

One obvious way to select actions at minimal computational costs is to use heuristics that do not require forward planning towards a goal (Gigerenzer & Gaissmaier, 2011; Soltani et al., 2016), e.g. to always select the action towards a hard to achieve and highly rewarded goal. Clearly, this and other heuristics come with the drawback that they can be substantially suboptimal when close to the goal. For example, blindly working toward a hard to achieve goal would ignore the risk of not reaching any goal. Another solution is to use habit-like strategies to avoid computational costs (Keramati et al., 2016). However, habits are typically useful only when one encounters exactly the same situation or context repeatedly, while goal reaching in uncertain environments as presented here, often requires flexible behavioural control.

It is an open question how humans select their actions when the potentially reachable goals are still far away and forward planning is complex. We hypothesized that people use a mixture of two approaches to achieve an acceptable balance between outcome and computational costs. This mixture changes with temporal distance to the goal: when far from the goal, people use a prior goal preference to make their decision about which action to take. With this approach, one assumes that one will eventually reach the preferred goal and selects the action that, if one looked backward in time from the reached goal, is the most instrumental. When coming closer to the goal, one expects that the influence of the goal preference should be progressively superseded by computationally more expensive action selection using forward planning to optimally reach the preferred goal or, failing that one, to pursue policies to reach an alternative goal.

To test whether participants used such an approach, we employed a novel behavioural task where participants were placed in a dynamic and stochastic sequential decision task environment that emulated reaching goals over an extended time period. In miniblocks of 15 trials, participants had to make decisions to reach one or two goals, where reaching both goals was rewarded more than reaching only one. In each miniblock, it was also possible, if blindly trying to obtain the higher reward, to not reach any goal and not obtain any reward. While participants pass through the miniblock, both the remaining trials to the end of the miniblock and the complexity of forward planning decrease. This enables us to test and model whether participants switch from using heuristics to forward planning during goal-reaching. To analyse the behavioural data of 89 participants and test hypotheses, we used stochastic variational inference, which provided posterior beliefs about the goal strategy preference of each participant, among other free model parameters. We show that the heuristic goal strategy preference parameter is key to explain participants' choices when temporally distant from the goal, and how, when progressing towards a goal, this goal strategy preference interacts with optimal forward planning to achieve near-optimal performance.

2.4 Methods

2.4.1 Participants

Eighty-nine participants took part in the experiment (58 women, mean age = 24.8, SD = 7.1). Reimbursement was a fixed amount of 8€ or class credit plus a performance-dependent bonus (mean bonus = 3.88€, SD = 13.6). The study was approved by the Institutional Review Board of the Technische Universität Dresden and conducted in accordance to ethical standards of the Declaration of Helsinki. All participants were informed about the purpose and the procedure of the study and gave written informed consent prior to the experiment. All participants had normal or corrected-to-normal vision.

Table 2.1. Glossary of abbreviations

Abbreviation	Explanation
A, B	Basic offers
Ab, aB	Mixed offers
Pts_t^A	A-points in trial t
Pts_t^B	B-points in trial t
g1	One-goal-choice = Sequential strategy choice = Choice that maximizes point difference
g2	Two-goal-choice = Parallel strategy choice = Choice that minimizes point difference
G1	One-goal-success = One point scale above threshold after 15 trials
G2	Two-goal-success = Both scales above threshold after 15 trials
Q(s,a)	Action value = Expected future reward of a choice
Q_G(s,a)	Goal choice value = Expected future reward of a goal strategy choice
DEV	Differential expected value = Q_G (s, g2) - Q_G (s, g1)

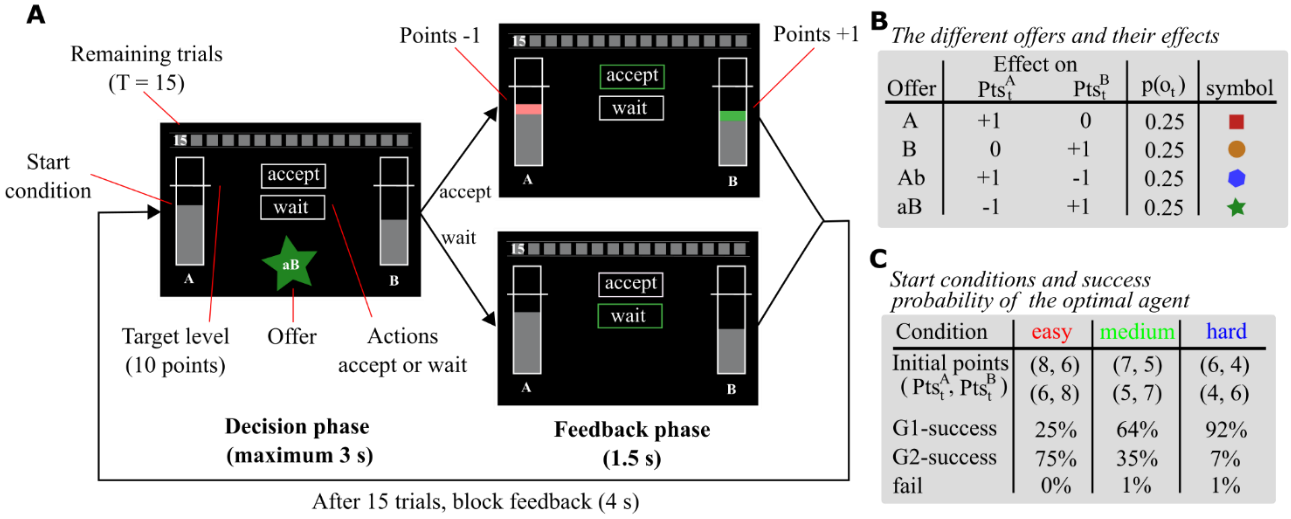
2.4.2 Experimental Task

The experiment included a training phase of 10 miniblocks, followed by the main experiment comprising 60 miniblocks. The 60 miniblocks in the main experiment were subdivided into three sessions of 20 miniblocks between which participants could make a self-determined pause. A miniblock consisted of $T = 15$ trials in which participants had to accept or reject presented offers to collect A-points (Pts_t^A) and B-points (Pts_t^B, see Table 2.1 for a glossary of abbreviations). If participants reached the threshold of 10 points for either A- or B-point scale after 15 trials, they received a reward of 5 cents. If participants reached the threshold for both point scales, they received a reward of 10 cents. If none of the two thresholds was reached, no additional reward was provided. In total, each participant completed 150 training trials and 900 trials in the main experiment.

Each trial started with a response phase lasting until a response was made, but not more than 3 seconds (Fig 2.1, A). The current amount of A-points and B-points was visualized by two vertical bars flanking the stimulus display. Horizontal white lines marked the threshold of 10 points. At the top of the screen, a grey timeline informed the participants about the remaining trials in the miniblock. The current offer was displayed at the bottom centre, and the two choice options were presented in the centre of the screen by the framed words ‘accept’ and ‘wait’. Participants could accept an offer by an upwards keypress and reject the offer by a downwards keypress. If participants did not respond within 3 seconds the trial was aborted, and a message was displayed reminding the participant to pay attention. If participants missed the response deadline more than 5 times in the whole main experiment, 50 cents were subtracted from their final payoff (mean number of timeouts = 1.34, SD = 1.7). After the response phase, feedback was displayed for 1.5 seconds. Response feedback included a change in colour of the frame around the selected response from white to green. Additionally, the gain or loss of points was visualized by colouring the respective area on the bar either green or red. After 15 trials, feedback for the miniblock was displayed for 4 seconds informing the participants whether they won 5, 10 or 0 cents. Code for experimental control and stimulus presentation was custom written in Matlab (MathWorks) with extensions from the Psychophysics toolbox (Kleiner et al., 2007).

Participants were presented with four different offers (A, B, Ab, and aB) that occurred with equal probability on each trial of the miniblock (see Fig 2.1, B). We call A or B basic offers and Ab or aB mixed offers. Accepting basic offers increased the corresponding point count, whereas accepting mixed offers transferred a single point from one scale to the other. The basic offers introduce a stochastic base rate of points, which allows participants to accumulate enough points on one or both point scales. In contrast, mixed offers allow us to identify participants’ intention to reach a state in which either both point scales are above threshold ($Pts_T^A \geq 10$ and $Pts_T^B \geq 10$) or only one point scale is above threshold (e.g. $Pts_T^A < 10$ and $Pts_T^B \geq 10;$ see below for more details). Rejecting an offer did not have any effect on the current point count. All participants received the same sequence of offers. We generated pseudorandomized lists for the training phase and for the three main experimental phases such that the frequency of offers reflected an equal offer occurrence probability in every list. We associated each offer with a coloured symbol to facilitate fast recognition.

Three different conditions modulated the difficulty to reach both thresholds by varying the number of initial points (Fig 2.1, C). We chose the number of initial points such that an optimal agent’s probability of reaching both thresholds was 75% in easy, 35% in medium and 7% in hard. The agent’s goal reaching performance for each initial point configuration was based on 10,000 simulated miniblocks with uniform offer probability (see below how we define the optimal agent). The same sequence of start conditions was presented to all participants. Pseudorandomized lists with a balanced frequency of initial point configurations were generated for the training phase and for the three main experimental phases. Note that the observed agent behaviour in the results section deviates from what we expected based on the experimental parametrization process. These discrepancies arise because we used random offer sequences (offers with equal probability) for experimental parametrization, but one specific offer sequence for the actual experiment. For example, in some miniblocks there were only few basic offers (see S2.1-2.4 Fig for details about the used offer sequence).



The different offers and their effects

Offer	Effect on Pts_t^A	Pts_t^B	$p(o_t)$	symbol
A	+1	0	0.25	🟥
B	0	+1	0.25	●
Ab	+1	-1	0.25	⬢
aB	-1	+1	0.25	★

Start conditions and success probability of the optimal agent

Condition	easy	medium	hard
Initial points (Pts_t^A, Pts_t^B)	(8, 6) (6, 8)	(7, 5) (5, 7)	(6, 4) (4, 6)
G1-success	25%	64%	92%
G2-success	75%	35%	7%
fail	0%	1%	1%

Fig 2.1. Experimental task. (A) Depiction of trial timeline and stimulus features. Participants performed miniblocks of 15 trials in which they collected points to reach either one or two goals, rewarding them with additional 5 or 10 Cents. Each trial started with a decision phase (maximum 3 seconds) in which participants had to accept or reject a presented offer. Depending on the offer, accepting increased or decreased A- and B-points. The current amount of points was displayed by two grey bars flanking the stimulus screen. In the feedback phase (1.5 seconds), gained points were displayed as a green area and lost points as a red area on the bar. The horizontal lines crossing the bars indicated the threshold for reaching goal A and goal B. After 15 trials, feedback for the miniblock was displayed (4 seconds) informing the participant about the reward gained. **(B)** Summary of offer types and their effect on point count**.** Offers occurred with equal probability in each trial of the miniblock. Basic offers (A and B) increased either A or B points. Mixed offers (Ab and aB) added one point on one side but subtracted one point on the other side. Only accepting an offer had an effect on points. **(C)** Three different conditions modulated the difficulty to reach both thresholds by varying the number of initial points. Using an optimal agent, we chose the number of initial points, such that the agent's probability of reaching both thresholds (G2-success) was 75% in easy, 35% in medium and 7% in hard.

2.4.3 Choice classification

In order to maximize reward, it was key for the participants to decide whether they should pursue the A- and B-goal in a sequential or in a parallel manner. A parallel strategy, i.e. balancing the two point scales, increases the likelihood that both goals (G2, see Table 2.1) will be reached at the end of the miniblock, but at the risk of failing. A sequential strategy, i.e. first secure one goal, then focus on the second one, might increase the likelihood to reach at least one goal (G1) within 15 trials, but decreases the likelihood to achieve G2.

To obtain a trial-wise measure of the pursued goal strategy, choices were classified based on the current point difference and the offer. Choices that minimized the difference between points were classified as two-goal-choice ($a_t = g2$), reflecting the intention to fill both bars using a parallel strategy. Choices that maximized the difference between points were classified as one-goal-choice ($a_t = g1$), reflecting the intention to pursue G1, or the intention to maintain one bar above threshold if G1-success has already been attained (see S2.1 Table). For example, if a participant has 8 A-points and 6 B-points and the current offer is Ab, accepting would be a g1-choice, whereas waiting would be a g2-choice. Conversely, for an aB offer, accepting would be a g2-choice and waiting a g1-choice. If the difference between points ($Pts_t^A - Pts_t^B$) is 1 and the offer is aB, g-choice is not defined because the absolute point difference would not be changed. This also applies to the mirrored case, where the difference between points ($Pts_t^A - Pts_t^B$) is -1 and the offer is Ab. Note that, due to the experimental design, response (accept/wait) and g-choice (g2/g1) were weakly correlated (r = 0.21). Furthermore, g-choice classification is only defined for the mixed offers (Ab and aB). The basic offers (A and B) are not informative with respect to the participants' pursued goal strategy. Importantly, all trial-level analysis will be restricted to trials which can be related to g-choices.

2.4.4 Task model

Here we will formulate the task in an explicit mathematical form, which will help us clarify what implicit assumptions we make in the behavioural model (Ostwald et al., 2018). We define a miniblock of the two-goal task as a tuple

$$(T, S, O, R, A, p(s_{t+1}|s_t, o_t, a_t), p(o_t), p(r_t|s_t)) \tag{2.1}$$

where

- $T = 15$ denotes the number of trials in a miniblock, hence $t = 1, \dots, 15$.
- $S = \{0, \dots, 20\}^2$ denotes the set of task states, corresponding to the point scale of the two point types (A, and B). Hence, a state s_t in trial t is defined as a tuple consisting of point counts along the two scales, $s_t = (Pts_t^A, Pts_t^B)$.

- $O = \{A, B, Ab, Ba\}$ denotes the set of four offer types, where the upper case letters denote an increase in points of a specific type and the lower case letters subtraction of points.
- $R = \{R_0, R_L, R_H\} = (0, 5, 10)$ denotes the set of rewards.
- $A = \{0, 1\}$ denotes the set of choices, where 0 corresponds to rejecting an offer and 1 to accepting an offer.
- $p(s_{t+1}|s_t, o_t, a_t)$ denotes state transitions which are implemented in a deterministic manner as $s_{t+1} = s_t + a_t * m(o_t)$, where $m(o_t)$ maps offer types into the point changes on the two point scales.
- $p(o_t = i) = \frac{1}{4}$ (for $\forall\ i \in O$) denotes a uniform distribution from which the offers are sampled.
- $p(r_t|s_t)$ denotes the state and trial dependent reward distribution defined as

$$p(r_t = R_0|s_t) = 1, \text{for } \forall t < T$$
$$p\big(r_T = R_L\big|Pts_T^A \geq 10\ \oplus\ Pts_T^B \geq 10\big) = 1$$
$$p\big(r_T = R_H\big|Pts_T^A \geq 10\ \wedge\ Pts_T^B \geq 10\big) = 1$$

Note that in the experiment the participants are exposed to a pseudo-random sequence of offers, meaning that within one experimental block all participants observed the same sequence of offers pre-sampled from this uniform distribution (see S2.1-2.4 Fig. for additional information about the used offer sequence). For simulations and parameter estimates we use the same pseudo-random sequence of observations, hence in each trial t of a specific block b offers are selected from a predefined sequence $o_{1:T}^{1:B} = (o_1^1, \dots, o_T^1, \dots, o_1^B, \dots, o_T^B)$, initially generated from a uniform distribution.

2.4.5 Behavioural model

To build a behavioural model, we assume that participants have learned the task representation through the training session and initial instruction. Hence, the behavioural model is represented by the following tuple

$$\big(T, S, O, R_\kappa, A, p(s_{t+1}|s_t, o_t, a_t), p(o_t), p(r_t|s_t)\big) \qquad (2.2)$$

where

- $T, S, O, A, p(s_{t+1}|s_t, o_t, a_t), p(o_t), p(r_t|s_t)$ are defined the same way as in the task model.
- $R_\kappa = \{0, 5, 10 \cdot \kappa\}$ denotes an agent-specific valuation of the rewarding states. Although the instructions for the experimental task clearly explained that participants receive a specific monetary reward depending on the final state reached during a miniblock, we considered a potential biased estimate of the ratio between G2 and G1 monetary rewards, quantified

with the free model parameter $\kappa \in [0, 2]$. In other words, we assumed that the participants might overestimate or underestimate the value of a G2-success, relative to a G1-success.

Importantly, the process of action selection corresponds to following a behavioural policy that maximises expected value during a single miniblock. We classified as G2-success miniblocks in which both point scales were above threshold after the final trial ($Pts_T^A \geq 10$ and $Pts_T^B \geq 10$). We classified as G1-success miniblocks in which only one point scale was above threshold (e.g. $Pts_T^A < 10$ or $Pts_T^B \geq 10$).

In what follows we derive the process of estimating choice values and subsequent choices based on dynamic programming applied to a finite horizon Markov decision process (Puterman, 2014). For experimental studies see also (Ballard et al., 2016; Korn & Bach, 2018).

2.4.6 Forward Planning

We start with a typical assumption used in reinforcement learning, namely that participants choose actions with the goal to maximize future reward. Starting from some state s_t at trial t, offer o_t, and following a behavioural policy π we define an expected future reward as

$$V[s_t, o_t|\pi] = \sum_{k=t+1}^{T} \gamma^{k-t-1} E[r_k|s_t, o_t, \pi] \tag{2.3}$$

where γ denotes a discount rate and $E[r_k|\, s_t, o_t, \pi]$ denotes expected reward at some future time step k. The behavioural policy sets the state-action probability $\pi(a_t, \dots, a_T|s_t, \dots, s_{T-1})$ over the current and future trials. Hence, we can obtain the expected reward as

$$E[r_k|s_t, \pi] = \sum_{r_k} r_k p(r_k|s_t, \pi) \tag{2.4}$$

where

$$p(r_k|s_t, \pi) = \sum_{s_{t+1:k}} \sum_{a_{t:k-1}} p(r_k|s_k) \prod_{\tau=t+1}^{k} p(s_\tau|s_{\tau-1}, o_{\tau-1}, a_{\tau-1})\, p(o_{\tau-1}) \pi(a_{\tau-1}|s_{\tau-1}\,) \tag{2.5}$$

Note that we use $s_{t+1:k}$, and $a_{t:k-1}$ to denote a tuple of sequential variables, hence $x_{m:n} = (x_m, \dots, x_n)$. The key step in deriving the behavioural model was to find the policy which maximises the expected future reward, that is, the expected state-offer value. In practice, one obtains the optimal policy as

$$\pi^* = \underset{\pi}{\operatorname{argmax}} V[s_t, o_t|\pi] \tag{2.6}$$

We solve the above optimization problem using the backward induction method of dynamic programming. The backward induction algorithm is defined in the following iterative steps:

(i) set the value of final state s_T as the reward obtained in that state $V[s_T|\pi^*] = \sum_{r_T \in R\_K} r_T \, p(r_T|s_T)$

(ii) compute state-offer-action value as $Q(s_k, o_k, a_k) = \gamma \sum_{s_{k+1}} V[s_{k+1}|\pi^*] p(s_{k+1}|s_k, o_k, a_k)$

(iii) set optimal choice for given state-offer pair as $a_k^* = \underset{a}{\operatorname{argmax}} Q(s_k, o_k, a)$

(iv) define the expected value of state s_k under optimal policy π^* as $V[s_k|\pi^*] = \sum_{o_k} Q(s_k, o_k, a_k^*) p(o_k)$

(v) repeat steps (ii) – (iv) until $k = t$

Hence, for a fixed value of the reward ratio (κ) an optimal choice at trial t corresponds to

$$a_t^* = \underset{a}{\operatorname{argmax}} Q(s_t, o_t, a) \tag{2.7}$$

We will define the optimal agent as an agent who has a correct representation of the reward ratio ($\kappa = 1$) and does not discount future reward ($\gamma = 1$). We illustrate in Fig 2.2 the Q-value to accept, estimated for the case of the optimal agent in an example trial ($Pts_t^A = 8$, $Pts_t^B = 11$, $o_t = Ab$).

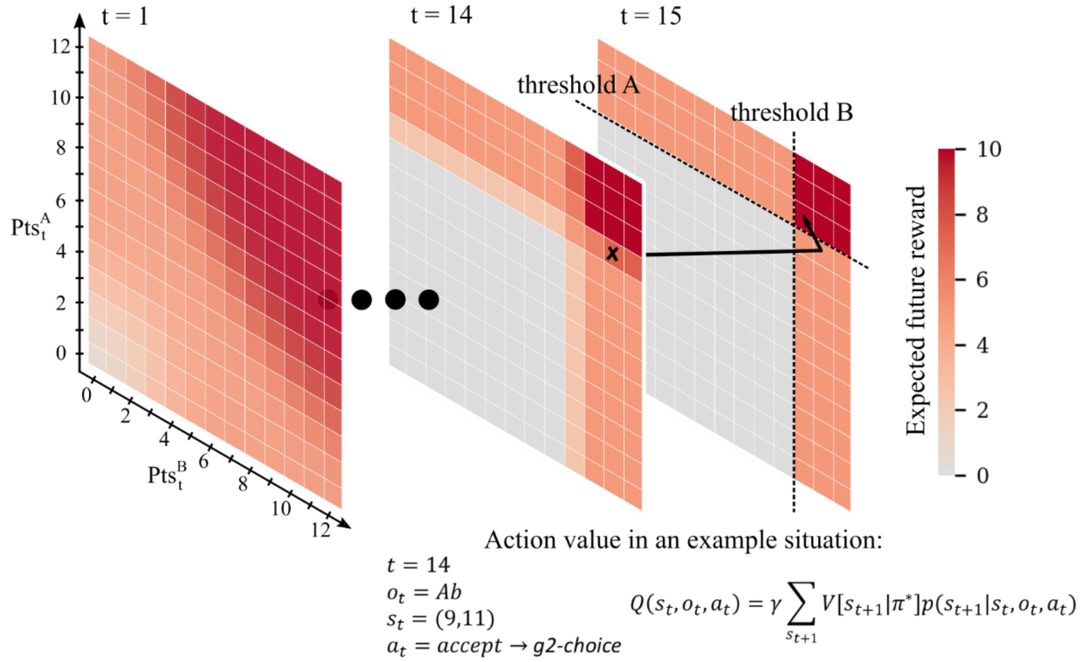


Fig 2.2. Illustration of the state space and associated expected future reward for the optimal agent (γ = 1, κ = 1). The black arrow shows a hypothetical transition in the state space. In trial 14 the participant has 9 A-points and 11 B-points (marked by the black cross) and accepts an offer Ab, gaining one A-point and losing one B-point (g2-choice). In the resulting state, both thresholds are reached; thus, the value of that state is 10 Cents. Similarly, the action that leads to that state has an associated Q-value of 10 Cents. In this example the agent would just have to wait in the last trial (15) to gain a 10 cents reward.

2.4.7 Response likelihood

Participants might compute expected values by mentally simulating and comparing sequences of actions towards the end of the miniblock. To illustrate the benefits of planning we consider the following example: There are 3 trials left in the current miniblock, and the participant has 9 A-points and 9 B-points (10 is threshold), and she receives offer Ab. Planning would, for example, allow to compute the probabilities for G2 when choosing either wait or accept. By waiting the participant would enter the second last trial with 9 A-points and 9 B-points. Receiving offer A or B in the second last trial (0.5 probability) followed by the complementary offer A or B in the last trial (0.25 probability) would grant G2. When choosing accept, the participant will have in the second last trial 10 A-points and 8 B-points. Consequently, she would need two consecutive B-offers (0.25 *0.25 probability) to achieve G2. Hence, by planning ahead one would conclude that wait gives the highest probability for a G2-success.

Still, planning an arbitrary number of future steps is complex and unrealistic. Hence, we make an assumption that the process of optimal action selection described above is perturbed by noise (planning noise, and response noise) which we quantify in the form of a parameter β, denoting response precision. Hence, this precision parameter is critical to characterize the participants' reliance on forward planning. Furthermore, instead of an elaborate planning process participants might use a simpler heuristic when deciding which action to select. We capture this heuristic in form of an additional offer-state-action function $h(o_t, s_t, a_t, \theta)$ which evaluates choices relative to possible goals. We describe this heuristic evaluation below. Overall, we can express the response likelihood (the probability that a participant makes choice a_t) as

$$p(a_t|\beta,\theta,\gamma,\kappa) = s\big(\beta Q(o_t, s_t, a_t, \gamma, \kappa) + h(o_t, s_t, a_t, \theta)\big) \qquad (2.8)$$

where $s(x)$ denotes the softmax function.

2.4.8 Choice heuristic

The choice heuristic is defined relative to the current offer o_t, current state s_t, and possible choices a_t. Importantly, we will interpret the choice heuristic in terms of participants' biases towards approaching both goals in a sequential or parallel manner. Hence, it is more intuitive to define the choice heuristic as choice biases relative to the goals, and not accept-reject choices. The choice heuristic is defined as follows

$$h(o_t, s_t, a_t, \theta) = \begin{cases} \infty, for\ o_t \in \{A, B\}, and\ a_t = 1 \\ \theta, for, o_t \in \{Ab, Ba\}, and\ a_t \equiv g2 \\ 0, otherwise \end{cases} \tag{2.9}$$

where $a_t \equiv g2$ denotes choices (accept or reject) which can be classified as g2-choices (see subsection Choice classification for details). In summary, a choice which reduces the point difference ($Pts_t^A - Pts_t^B$), for the given offer and the current state, is classified as g2-choice and choice which increases the point difference as g1-choice. Essentially, the strategy preference parameter θ reflects participants' preference for pursuing a sequential (negative values) or parallel (positive values) strategy. For example, some participants might have a general tendency to pursue goals in a parallel manner, independent of the actual Q-values. Conversely, participants may prefer a more cautious sequential approach. Note that we expected this parameter to make the most significant contribution to participants' deviation from optimal behaviour, reflecting their reliance on decision heuristics early in the miniblock.

Finally, for those choices which can be classified as g2- or g1-choices, we can express the response likelihood in a simplified form, in terms of free model parameters $\beta, \theta, \gamma, \kappa$ (Table 2.2). We refer to the difference between Q-values for g-choice as the differential expected value (DEV),

$$DEV = \ Q_G(a_t = g2) - Q_G(a_t = g1) \tag{2.10}$$

Using DEV, we defined the probability of making a g2-choice as

$$p(g2) = \sigma(\beta \cdot DEV(\gamma, \kappa) + \theta) \tag{2.11}$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ denotes the logistic function. Note that the probability of g1-choice becomes $p(g1) = 1 - p(g2)$.

Table 2.2. Summary of four free model parameters, the variables, the transformations used to map values to unconstrained space and their function in modelling participant behaviour.

Name	Variable	Transform	Function
Precision	β	$x_1 = \ln \beta$	Captures the impact of DEV, derived by forward planning, on action selection
Strategy preference	θ	$x_2 = \theta$	Heuristic preference of pursuing a parallel ($\theta > 0$) or sequential ($\theta < 0$) strategy, independent of the actual DEV
Discount rate	γ	$x_3 = \ln \frac{\gamma}{1-\gamma}$	Temporal discounting of DEV by the factor γ^{T-t}, where $T - t$ is the number of remaining trials
Reward ratio	κ	$x_4 = \ln \frac{\kappa}{2-\kappa}$	Accounts for the possibility that participants may overweight ($\kappa > 1$) or underweight ($\kappa < 1$) the actual reward for G2-success relative to G1-success.

2.4.9 Optimal agent comparison and general data analysis

We compared participant behaviour with simulated behaviour of an optimal agent. To summarize, we denote the optimal agent as the agent which has a correct representation of the reward function ($\kappa = 1$), does not discount future rewards ($\gamma = 1$), is not biased in favour of any choice ($\theta = 0$), and who generates deterministic g-choices based on DEV-values (corresponding to $\beta \rightarrow \infty$ in the response likelihood, that is, the argmax operator). The optimal agent deterministically accepts A and B offers.

When simulating agent behaviour to evaluate successful goal reaching, the agent received the same sequence of offers and initial conditions as the participants. Analysis on the level of g-choices was performed by registering instances in which the g-choice of a participant differed from the g-choice the optimal agent would have made in the same context (Pts_t^A, Pts_t^B, o_t, t). Trials with A or B offers and trials in which G2 had already been reached, were excluded from the g-choice analysis.

The goal of this comparison between summary measures of both optimal agent and participants was two-fold: First, we used this comparison to visualize deviations from optimality and motivate the model-based analysis which was used to test the hypothesis that a shift from heuristics to forward planning may explain these deviations. Second, plotting suboptimal g-choices instead of g-choices (Fig. Fig **2.4**) makes behaviour between participants more comparable. Plotting the proportion of g-choices averaged across participants would have been mostly uninformative because the significance of a g-choice depends on the current state, which is a consequence of the individual history of past choices within a miniblock. By registering deviations from an optimal reference point, we circumvent this state dependence of g-choices.

We used a sign test as implemented in the “sign_test” function of python’s “Statsmodels” (Seabold & Perktold, 2010) package to test whether participants total reward and success rates differed significantly from the optimal agent’s deterministic performance. We reported the p-value and the m-value $m = (N(+) - N(-))/2$, where $N(+)$ is the number of values above 0 and $N(-)$ is the number of values below and. To test for learning effects (in the main experimental phase), we used mixed effects models as implemented in R (Team, 2013) with the “lm4” package (Bates et al., 2014). Intercepts and slopes were allowed to vary between participants. p-values were obtained using the “lmerTest” package (Kuznetsova et al., 2017).

2.4.10 Hierarchical Bayesian data analysis

To estimate the free model parameters (Table 2.2) that best match the behaviour of each participant, we applied an approximate probabilistic inference scheme over a hierarchical parametric model, so-called stochastic variational inference (SVI) (Hoffman et al., 2013).

As a first step, we define a generic (weakly informative) hierarchical prior over unconstrained space of model parameters. In Table 2.2 we summarize the roles of free model parameters of our behavioural model and the corresponding transforms that we used to map parameters into an unconstrained space. We use $\boldsymbol{x}^n$ to denote a vector of free and unconstrained model parameters corresponding to the *n*th participant. Similarly, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ will denote hyperpriors over group mean and variance for each free model parameter. We can express the hierarchical prior in the following form

$$\mu_i \sim N(m_i, s_i) \tag{2.12}$$

$$\sigma_i \sim C^+(0,1) \tag{2.13}$$

$$x_i^n \sim N(\mu_i, \lambda\sigma_i) \tag{2.14}$$

$$\text{for } i \in [1, \dots, d], \text{and } n \in [1, .., N] \tag{2.15}$$

where $C^+(0,1)$ denotes a Half-Cauchy prior with scale $s = 1$, d number of parameters, and N number of participants. Note that by using this form of a hierarchical prior we make an explicit assumption that parameters defining the behaviour of each participant are centred on the same mean and share the same prior uncertainty. Hence, both the prior mean and uncertainty for each parameter are defined at the group level. Furthermore, the hyper-parameters of the prior $\eta = (m_1, \dots, m_4, s_1, \dots, s_4, \lambda)$ are also estimated from the data (Empirical Bayes procedure) in parallel to the posterior estimates of latent variables $\theta = (\mu_1, \dots, \mu_4, \sigma_1, \dots, \sigma_4, \boldsymbol{x}^1, \dots, \boldsymbol{x}^N)$. For more details, see supporting information (S2.1 Notebook).

The behavioural model introduced above defines the response likelihood, that is, the probability of observing measured responses when sampling responses from the model, condition on the set of model parameters $(\boldsymbol{x}^1, \dots, \boldsymbol{x}^N)$. The response likelihood can be simply expressed as a product of response probabilities over all measured responses $A = (\boldsymbol{a}^1, \dots, \boldsymbol{a}^N)$, presented offers $O = (\boldsymbol{o}^1, \dots, \boldsymbol{o}^N)$, and states (point configurations) visited by each participant $S = (\boldsymbol{s}^1, \dots, \boldsymbol{s}^N)$ over the whole experiment

$$p(A|O, S, \boldsymbol{x}^1, \dots, \boldsymbol{x}^N) = \prod_{n=1}^{N} \prod_{b=1}^{M} \prod_{t=1}^{T} p(a_{b,t}^n | s_{b,t}^n, o_{b,t}^n, \boldsymbol{x}^n) \tag{2.16}$$

where *b* denotes experimental block, and *t* a specific trial within the block.

To estimate the posterior distribution (per participant) over free model parameters, we applied the following approximation to the true posterior

$$p(\boldsymbol{x}^1, \dots, \boldsymbol{x}^N, \boldsymbol{\mu}, \boldsymbol{\sigma} |\, A, S, O) \approx Q(\boldsymbol{\mu}, \boldsymbol{\sigma}) \prod_{n}^{N} Q(\boldsymbol{x}^n) \tag{2.17}$$

$$Q(\boldsymbol{\mu}, \boldsymbol{\sigma}) = \frac{1}{\sigma_1 \dots \sigma_d} \mathcal{N}_{2d}\left(\boldsymbol{z};\ \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right) \text{ for } \boldsymbol{z} = (\mu_1, \dots, \mu_d, \ln \sigma_1, \dots, \ln \sigma_d) \tag{2.18}$$

$$Q(\boldsymbol{x}^n) = \mathcal{N}_d(\boldsymbol{x}^n;\ \boldsymbol{\mu}_x^n, \boldsymbol{\Sigma}_x^n) \tag{2.19}$$

Note that the approximate posterior captures posterior dependencies between free model parameters (in the true posterior) on both levels of the hierarchy using the multivariate normal and multivariate log-normal distributions. However, for practical reasons, we assume statistical independence between different levels of the hierarchy, and between participants. Independence between participants is justified by the structure of both response likelihood (responses are modelled as independent and identically distributed samples from conditional likelihood) and hierarchical prior (a priori statistical independence between model parameters for each participant).

Finally, to find the best approximation of the true posterior given the functional constraints of our approximate posterior, we minimized the variational free energy F[Q] with respect to the parameters of the approximate posterior.

$$-\ln p(A|S, 0) = F[Q] - D_{KL}(Q||p) \leq F[Q] = \mathrm{f}(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\mu}_x^1, \boldsymbol{\Sigma}_x^1, \dots, \boldsymbol{\mu}_x^N, \boldsymbol{\Sigma}_x^N) \tag{2.20}$$

$$F[Q] = \int \mathrm{d}\boldsymbol{x}^1 \dots \mathrm{d}\boldsymbol{x}^N \mathrm{d}\boldsymbol{\mu} \mathrm{d}\boldsymbol{\sigma} Q(\boldsymbol{\mu}, \boldsymbol{\sigma}) \prod_{n}^{N} Q(\boldsymbol{x}^n) \ln \frac{Q(\boldsymbol{\mu}, \boldsymbol{\sigma}) \prod_n^N Q(\boldsymbol{x}^n)}{p(A|O, S, \boldsymbol{x}^1, \dots, \boldsymbol{x}^N) p(\boldsymbol{x}^1, \dots, \boldsymbol{x}^N, \boldsymbol{\mu}, \boldsymbol{\sigma})} \tag{2.21}$$

The optimization of the variational free energy F[Q] is based on the SVI implemented in the probabilistic programming language Pyro (Bingham et al., 2018) and the automatic differentiation module of PyTorch (Paszke et al., 2017), an open source deep learning platform.

As a final remark, we would like to point out that it is possible to use a different hierarchical prior (Polson & Scott, 2010), different parametrization of the hierarchical model (Bernardo et al., 2003) or different factorization of the approximate posterior (e.g., mean-field approximation). However, through extensive comparison of posterior estimates on simulated data, we have determined that the presented hierarchical model and the corresponding approximate posterior provide the best posterior estimate of free model parameters among the set of parametric models we tested (S2.1 Notebook).

2.5 Results

To investigate how the balance between computationally costly forward planning and heuristic preferences changes as a function of temporal distance from the goals, participants performed sequences of actions in a novel sequential decision making task. The task employed a two-goal setting, where participants had to decide between approaching the two goals in a sequential or in a parallel manner. We first performed a standard behavioural analysis, followed by a model-based approach showing that participants use a mixture of strategy preference and forward planning to select their action.

2.5.1 Standard behavioural analysis

We first analysed the general performance of all participants and – for each miniblock and trial – compared it to the behaviour of an optimal agent possessing perfect knowledge of the task and performing full forward planning to derive an optimal policy that maximizes total reward. The motivation of this comparison was to detect differences between how the optimal agent and participants perform the task. These differences will motivate our model-based analysis below. To compute and compare optimal vs individual policies, all participants and the agent received exactly the same sequence of offers and start conditions. The difference in total reward between participants and agent was significant (m = -35.5, $p < 0.001$), where participants earned 388.5 Cents (SD = 13.6) and the agent earned 405 Cents. As expected, both participants and agent earned more money in the easy condition than in the medium condition and least in the hard condition (Fig 2.3, A, C). In the easy and medium condition, the agent earned significantly more than the participants (easy: M = 8.7 Cents, SD = 8.4, m = -33, $p < 0.001$; medium: M = 7.2 Cents, SD = 7.0, m = -30, $p < 0.001$). In the hard condition, the total reward did not differ significantly between the participants and agent, m = 0.5, p > 0.99 (Fig 2.3, E). These results show that participant performance was generally close to the optimal agent but differed significantly in the easy and medium condition.

Next, we analysed participants' goal reaching success and compared it to the optimal agent. There were three possible outcomes in a miniblock: Achieving G1 (goal A or B), achieving G2 (A & B) or fail (neither A nor B). The main experiment comprised 20 miniblocks of each difficulty level modulating difficulty to reach G2. As expected, participants reached on average G2 more often in the easy (M = 71%, SD = 8%) than in the medium condition (M = 25%, SD = 6%), m = 44.5, $p < 0.001$. In the hard condition, participants reached G2 in only 1% (SD = 2%) of the miniblocks. Participants failed to reach any goal in 2% (SD = 3%) of the miniblocks in the medium and in 6 % (SD = 5%) of the miniblocks in the hard condition. They never failed in the easy condition (Fig 2.3, B). The agent reached G2 in 80% in the easy, in 30% in the medium and in 0% in the hard condition (Fig 2.3, D). Note that G2 cannot be reached in all miniblocks. We simulated all possible choice sequences (n = 2^15) for a given miniblock and evaluated whether G2 was theoretically possible. According to these simulations, 90% G2 performance can be reached in the easy, 35% in the medium and 5% in the hard condition.

When comparing participants' goal reaching success with the agent, we found that, on average, there was a consistent pattern of deviations in the easy and medium conditions (Fig 2.3, F). In the easy condition, participants reached G2 on average 9% (SD = 8%) less often than the agent (m = -33, $p < 0.001$), but reached G1 9% (SD = 8%) more often (m = 33, $p < 0.001$). In the medium condition, participants reached G2 on average 6% (SD = 6%) less often than the agent (m = -26, $p < 0.001$) but reached G1 4% (SD = 7%) more often (m = 16.5, $p < 0.001$). While the agent never failed, participants had a 2% (SD = 3%) fail rate (m = 11.5, $p < 0.001$). In the hard condition, participants reached G2 on average 0.6% (SD = 1.6%) more often than the agent (m = 5.5, $p < 0.001$). G1 (m = -7, p = 0.087) and fail-rate (m = 3.5, p = 0.42) did not differs significantly between participants and agent. In summary, these differences in successful goal reaching between participants and the agent explains the difference in accumulated total reward: Participants obtained less reward than the agent because on average they missed some of the opportunities to reach G2 in the easy and medium condition and sometimes even failed to achieve any goal in the medium and hard condition.

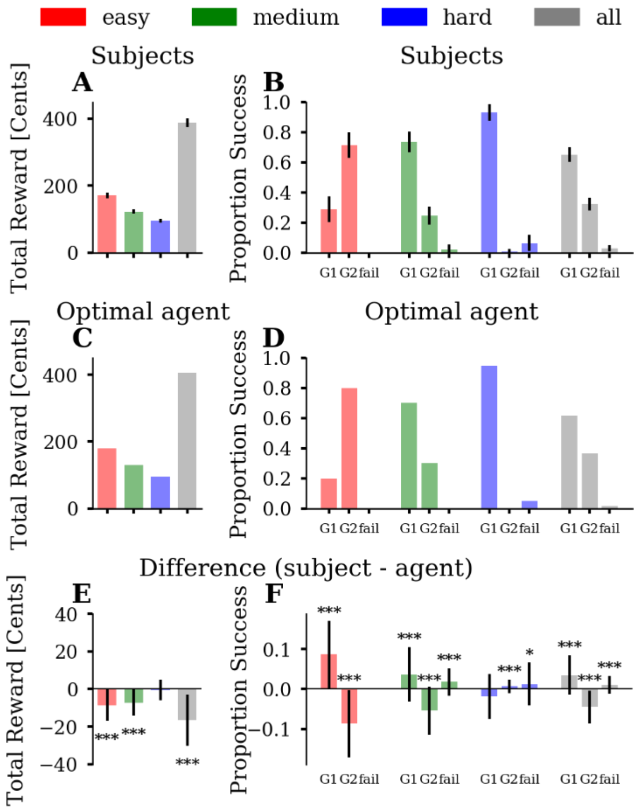


Fig 2.3. Standard analyses of total reward and comparison to the optimal agent. (A) Average total reward across participants. The three conditions are colour-coded (easy = red, medium = green, blue = hard) and the average over conditions is shown in grey. Error bars depict the standard deviation (SD). **(B)** Proportion of successful goal-reaching averaged across participants, for each of the three conditions. We plot the proportion of reaching, at the end of a miniblock, a single goal (G1), both goals (G2), or no goal (fail). The fourth block of bars in grey represents the proportions averaged over all three conditions. Error bars depict SD. **(C)** Simulated total reward of the optimal agent. **(D)** The goal-reaching proportions of the optimal agent. **(E)** Average difference between participants and agent with error bars depicting SD. **(F)** Averaged difference of proportion success between participants and agent with error bars depicting SD. One can see that the average goal-reaching proportions of participants were close to the agent's proportions. However, participants, on average, reached G2 less often than the agent. Asterisks indicate differences significantly greater than zero (Sign-test, * ≙ $p < 0.05$, ** ≙ $p < 0.01$, *** ≙ $p < 0.001$).

How can these differences in goal-reaching success be explained? To address this, we used the mixed-offer trials to identify which strategy a participant was pursuing in a given trial and compared the strategy choice to what the agent would have done in this trial. We classified strategy choices as evidence either of a parallel or a sequential strategy. With the parallel strategy (g2), participants make choices to pursue both goals in a parallel manner, while with a sequential strategy (g1), participants make choices to reach first a single goal and then the other. We inferred that participants used a g2-choice for a specific mixed-offer trial when the difference between the points of the two bars was minimized, while we inferred a g1-choice when the difference between points was maximized (see Methods). We categorized a participant's g2-choice as suboptimal when the optimal agent would have made a g1-choice in a specific trial and vice versa. Fig 2.4, A-D shows the proportions of suboptimal g-choices in mixed-offer trials. In the easy condition, participants made barely any suboptimal g2-choice (mean = 0%, SD = 0.001%), but 29% (SD = 10%) suboptimal g1-choices (Fig 2.4, A). This means that participants, on average, preferred a sequential strategy more often than would have been optimal. In the medium condition participants made on average 6% (SD = 3%) suboptimal g2-choices and 28% (SD = 11%) suboptimal g1-choices. Similar to the easy condition, participants, on average, preferred a sequential strategy where a parallel strategy would have been optimal. In the hard condition, this pattern reversed. Participants made on average 40% (SD = 12%) suboptimal g2-choices, relative to the agent, and 11% (SD = 6%) suboptimal g1-choices. Participants' suboptimal g-choices were also reflected in goal reaching success. In the easy and medium condition, suboptimal g1-choices, relative to the agent, resulted in a higher proportion of reaching G1, and a lower proportion of reaching G2. In the hard condition, suboptimal g2-choices led to occasional fails and a tiny margin of reaching G2. However, despite suboptimal g2-choices, participants still reached G1 in 93% (SD = 6%) of the miniblocks.

As the first test of our prediction that participants tend to use more forward planning when temporally proximal to the goal, we analysed suboptimal decisions as a function of trial time. As expected, suboptimal decisions, relative to the agent, decreased over trial time (Fig 2.4, B). While in the first trial, 42% (SD = 19%) of participants' g-choices deviated from the agent's g-choices, participant behaviour converged to almost optimal performance towards the end of the miniblock, with only 4% deviating g-choices (SD = 7%). We also simulated a random agent that accepts all basic A or B offers but guesses on mixed offers (S6-7 Fig). S2.7 Fig B shows that the random agent makes approximately 50 % suboptimal g-choices across all trials in the miniblock. That means participants used non-random response strategies, i.e. planning or heuristics, since their pattern of suboptimality across trials deviated from the straight-line pattern of the random agent.

In the hard condition, the number of suboptimal g2-choices similarly decreased, but not in the easy and medium condition (Fig 2.4, C). The number of suboptimal g1-choices decreased across trials in the easy and medium, but not in hard condition (Fig 2.4, D). Note that in easy and the medium conditions, opportunities to make suboptimal g2-choices are generally scarce, because the difference between action values $DEV = Q_G(g2) - Q_G(g1)$ was mostly positive, which means that a g2-choice was mostly optimal. Similarly, in the hard condition, as there was a low number of opportunities to make suboptimal g1-choices, there was no clear decrease in the number of suboptimal g1-choices.

Although these findings of diminishing suboptimal choices over the course of miniblocks may be explained by the participants' initial employment of a suboptimal heuristic, there is an alternative explanation because we used an optimal agent, which uses a max operator to select its action: If this agent computes, by using forward planning, a tiny advantage in expected reward of one action over the other, the agent will always choose in a deterministic fashion the action with the slightly higher expected reward. Therefore, at the beginning of the miniblock, where the distance to the final trial is largest, the difference between goal choice values $DEV = Q_G(g2) - Q_G(g1)$ (S2.5 Fig) is close to 0. The reason for this is that a single g2-choice at the beginning of the miniblock does not increase the probability for G2-success by much. However, when only few trials are left, a single g2-choice might make the difference between winning or losing G2. Since $DEVs$ are close to 0 at the initial trials we cannot exclude the possibility yet that participants actually may have used optimal forward planning just like the agent but did not use a max operator. Instead, participants may have sampled an action according to the computed probabilities of each action to reach the greater reward in the final trial. Such a sampling procedure to select actions would also explain the observed pattern of diminishing suboptimal g-choices over the miniblock (Fig. Fig **2.4** B-C). To answer the question, whether there is actually evidence that participants use heuristics, when far from the goal, even in the presence of probabilistic action selection of participants, we now turn to a model-based analysis.

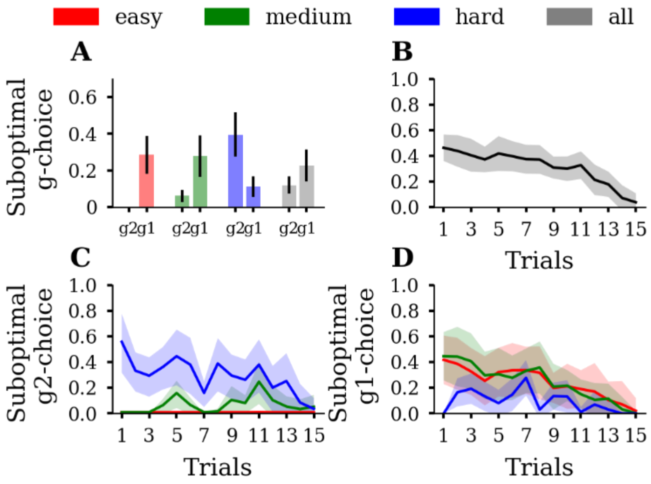


Fig 2.4. Suboptimal choices. (A) Proportions of suboptimal g1-choices (g1) and suboptimal g2-choices (g2), averaged over participants. Participants tend to make suboptimal g1-choices in the easy and medium condition while this pattern reverses in the hard condition. Error bars depict SD. Conditions are colour coded. **(B)** Suboptimal g-choices as a function of trial averaged over participants. Shaded areas depict SD. **(C)** Suboptimal g2-choices as a function of trial averaged over

participants. **(D)** Suboptimal g1-choices as a function of trial averaged over participants. In both C and D, one can see that participants made more suboptimal g-choices at the beginning of the miniblock than close to the final trial. Shaded areas depict SD.

2.5.2 Model-based behavioural analysis

To infer the contributions of participants' forward planning and heuristic preferences, we conducted a model-based analysis. If we find that participants' strategy preference θ is smaller or larger than zero, we can conclude that participants indeed used a heuristic component to complement any forward planning. This is especially relevant for choices early in the miniblock as DEV values are typically close to zero. Indeed, when inferring the four parameters for all 89 participants using hierarchical Bayesian inference, we found that participants' g-choices were influenced by a heuristic strategy preference in addition to a forward planning component (Fig 2.5, A). For 74 out of 89 participants, we found that the 90% credibility interval (CI) of the posterior over strategy preference did not include zero. 68 of these participants had a positive strategy preference, meaning they preferred an overall strategy of pursuing both goals in parallel. Six of these participants had a negative strategy preference, meaning they preferred to pursue both goals sequentially. The median group hyperparameter of strategy preference was 0.55 (90% CI = [0.47, 0.63]). For example, a participant with this median strategy preference, in a mixed-offer trial where $DEV = 0$, would make a g2-choice with 63% probability, whereas a participant without a strategy preference bias, i.e. $\theta = 0$, would make a g2-choice with 50% probability. After the experiment, we had asked participants whether they used any specific strategies to solve the task and to give a verbal description of the used strategy. Reports reflected three main patterns: Pursuing one goal after the other (sequential strategy), promoting both goals in a balanced way (parallel strategy), and switching between sequential and parallel strategy, depending on context (mixed strategy). Reported strategies are in good qualitative agreement with the estimated strategy preference parameter (S2.8 Fig), supporting our interpretation of this parameter. Notably, the task instructions, given to the participants prior to the experiment, did not point to any specific heuristic (S2.1 Text). Altogether, the non-zero strategy preference in 83% of participants indicates that suboptimal decisions within a miniblock (see Fig 2.4) are not only caused by probabilistic sampling for action selection, but also by the use of a heuristic strategy preference.

As expected, we found that the DEV (see Table 2.1) derived by forward planning influenced action selection (median group hyperparameter of the inferred precision β = 1.82, 90% CI = [1.45, 2.3], Fig 2.5, B). For example, a hypothetical participant with parameters similar to the group hyperparameters ($\theta = 0.55$ and $\beta = 1.82$), when encountering a $DEV = 0.5$, would make a g2-choice with 82% probability. Increasing DEV by 1 would increase the g2-choice probability to 96%. In contrast, a participant with low precision but the same median strategy preference ($\theta = 0.55$ and $\beta = 0.5$), when encountering a $DEV = 0.5$, would make a g2-choice with 69% probability. Increasing DEV by 1 would increase g2-choice probability to 79%. We found evidence only for weak discounting of future rewards, as for most participants the inferred discount was close to 1 (median of the inferred discount parameter γ = 0.984, 90% CI = [0.978, 0.988], Fig 2.5, C). We found that some participants used a reward ratio different from the objective value of 1 (CI not containing 1). Twelve participants had a reward ratio greater than 1 and 17 participants had a reward ratio smaller than 1. However, the median group hyperparameter of the inferred reward ratio was close to the objective value of 1 (κ = 1.05, 90% CI = [0.99, 1.11], Fig 2.5, D). A reward ratio of 1.2 means, that participants behaved as if the value of achieving G2 would be 2.4 times the value of achieving G1(when in reality the reward is only double as high). While strategy preference has its greatest influence during the first few trials of a miniblock, the reward ratio has an influence only when forward planning, i.e. changes the DEV, and will therefore affect action selection most during the final trials of a miniblock. In addition, we found only low posterior correlation between the strategy preference and reward ratio parameter, indicating that these two parameters model distinct influences on goal reaching behaviour.

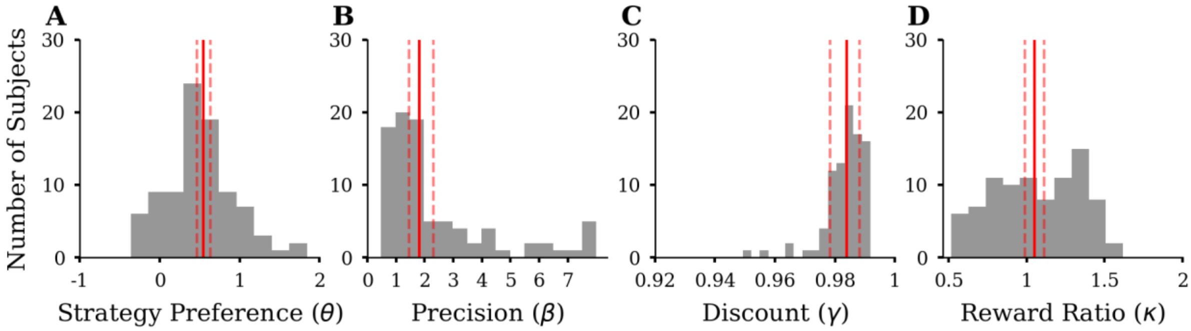


Fig 2.5. Summary of inferred parameters of the four-parameter model for all 89 participants. We show histograms of the median of the posterior distribution, for each participant. Solid red lines indicate the median of the group hyperparameter posterior estimate with dashed lines indicating 90% credibility intervals (CI). **(A)** Histogram of strategy preference parameter θ. **(B)** Histogram of precision parameter β (last bin containing values > 8). **(C)** Histogram of discount parameter γ. **(D)** Histogram of reward ratio parameter κ.

To show that our model with constant parameters is able to capture a dynamic shift from heuristic decision making to forward planning we conducted two sets of simulations where we systematically varied the response precision β and the strategy preference parameter θ. First, we simulated behaviour where we varied β between 0.25 and 3 with θ, γ, and κ sampled from their fitted population mean (S2.1-2.2 Movie). S2.2 Movie, B shows that the higher β, the fewer suboptimal g-choices are made towards the end of the miniblock. Second, we simulated behaviour where we varied θ varied between -1 and 1 with β, γ, and κ sampled from their fitted population mean (S2.3-2.4 Movie). S2.4 Movie, B shows that a change in θ affects the number of suboptimal g-choices made at the beginning but not at the end of the miniblock. To understand these results, one has to consider that, due to the experimental design, the differential expected value (DEV) computed by forward planning is correlated with trial number (S2.5 Fig). The average of absolute $DEVs$ ($\gamma = 1, \kappa = 1$) was M = 0.12 (SD = 0.12) in the first third, M = 0.29 (SD = 0.29) in the second third and M = 0.89 (SD = 1.38) in the last third of the miniblock. Given these experimental constraints, it becomes apparent that the fitted group hyperparameters $\theta = 0.55$ and $\beta = 1.82$ suggest, that participants' behaviour is best explained by a shift from heuristic decision making to forward planning. For small $DEVs$, the influence of the fitted β on choice probability is marginal; therefore, the relative influence of the fitted strategy preference parameter θ is high, and behaviour is driven by heuristic choices. For higher trial numbers, i.e. closer to the end of the miniblock, $DEVs$ tend to be high so that the model-based value ($\beta * DEV$) is large relative to the strategy preference θ; therefore, towards the end of the miniblock behaviour is driven by forward planning, with a transition from one decision mode to another in between. If participants would have planned ahead already in early trials, this would have been reflected in a large precision parameter ($\beta \gg \theta$), since small $DEVs$ in early trials, multiplied by large β, could dominate any heuristic bias. We also implemented a model with changing parameters over trials and compared it to the constant model. Parameters were fit separately for three partitions of the miniblock, i.e. early (trials 1- 5), middle (trials 6-10) and late trials (11-15). Model comparisons showed that this model with changing parameters had lower model evidence compared to the model with constant parameters (S2.9 Fig).We interpret these results as further evidence that the described constant parameterization is sufficient to describe a hidden shift from using a heuristics to forward planning.

Finally, as an additional test of the hypothesis that participants rely more on heuristic preferences when the goal is temporally distant, we conducted a multiple regression analysis (Fig 2.6, A). To do this, we divided the data into the first (first 7 trials) and the second half (last 8 trials) of miniblocks, and computed, for each participant the proportion of g2-choices in the mixed-offer trials. We fitted, across participants, these proportions of g2-choices against 6 regressors: strategy preference, precision, discount rate, reward ratio, a dummy variable coding for the first and second miniblock half and interaction between strategy preference and miniblock half. We found a significant interaction between strategy preference and miniblock-half (p < 0.001), demonstrating that strategy preference is more predictive for the proportion of g2-choices in the first half of the miniblock than in the second half. Fig 2.6, B visualizes the interaction effect showing that the slope of the marginal regression line for the first half of the miniblock is greater than the slope of the marginal regression line for the second half of the miniblock. This finding provides additional evidence that participants rely on heuristic preferences when the goal is temporally far away but use differential expected values (DEV) derived by forward planning when the goal is closer.

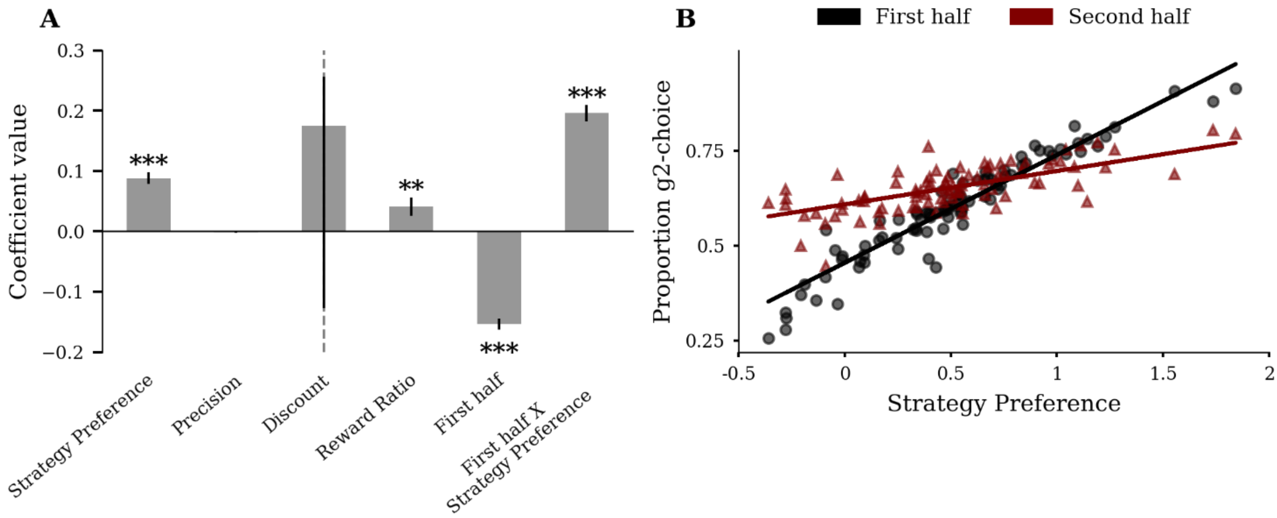


Fig 2.6. Strategy preference is more predictive for participant's proportion of g2-choices in the first than in the second half of the miniblock. (A) Linear regression of proportion g2-choice against parameters from the four-parameter model, a dummy variable coding for miniblock-half and interaction between miniblock-half and strategy preference. The significant interaction term supports the hypothesis that the influence of strategy preference on g2-choice proportion is greater in the first than in the second half of the miniblock. Error bars represent SE. Asterisks indicate coefficients significantly different from 0 (t-test, * ≙ p < 0.05, ** ≙ p < 0.01, *** ≙ p < 0.001). **(B)** Strategy preference plotted against the proportion of g2-choices in the first half of the miniblock (black) and in the second half of the miniblock (red). Solid lines represent marginal regression lines.

In addition, we conducted model comparisons, posterior predictive checks and parameter recovery simulations to test whether our model is an accurate and parsimonious fit to the data. First, we compared variants of our model, where we fixed individual parameters (S2.9 Fig). Adding θ and β increased model evidence, confirming their importance in explaining participant behaviour. The three-parameter model ($\theta,\ \beta,\ \kappa$) had the highest model evidence among all 16 models. Adding γ did not increase model evidence. This result is consistent since we found only little evidence for discounting when fitting the parameters, see Fig. Fig **2.5** C. To test whether participants used condition-specific response strategies (e.g., use heuristics in the easy and hard but plan forward in the medium difficult condition) we estimated model parameters separately for conditions. However, the condition-wise model had lower model evidence compared to the conjoint model, indicating that participants use a condition-general approach to arbitrate between using a heuristic and planning ahead. Second, we simulated data using the group mean parameters as inferred from the participants' data and compared it to the observed data. Visual inspection shows that both the simulated performance pattern (S2.10 Fig) and the simulated frequency of suboptimal g-choices (S2.11 Fig) closely resemble the experimentally observed patterns (Fig. 3 and 4). Third, we simulated data using participants' posterior mean and tested whether we could reliably infer parameters (S2.1 Notebook). Results showed that the inferred β, θ and κ align with the true parameter value, but simulation-based calibration (Talts et al., 2018) suggests that estimates of γ are biased. Taken together, our model provides a good fit to the data, where the data are informative about the three parameters β, θ and κ.

We also tested whether participants showed learning effects in the main experimental phase. In a first linear model, the depended variable was the total reward and the predictor was the experimental block number (miniblock 1-20, miniblock 21-40, miniblock 41-60). The analysis revealed a significant but small main effect of experiment block (β = 5.4, SE = 0.5, p < 0.001). In a second logistic model the dependent variable was suboptimal goal choice (1 = suboptimal, 0 = optimal) and the predictor was experiment block. The second analysis revealed a significant but small main effect of experiment block on the probability to make a suboptimal g-choice (β = -0.084, SE = 0.02, p < 0.001). Furthermore, we fitted the three parameter model (θ, β, κ) separately for experiment blocks. Model comparisons revealed that the experiment block-wise model had lower model evidence compared to the conjoint model (S2.9 Fig.).

As a final control analysis, we used logistic regression to establish how the absolute difference between A- and B-points affects goal choice as a function of the number of trials remaining in the miniblock. If participants rely on a fixed strategy preference when far from the goal, there should be no effect of absolute score difference on goal choice at the start of miniblocks. In this model the depended variable was goal choice (1 = g2, 0 = g1) and the predictors were absolute score difference ($|Pts_t^A - Pts_t^B| \in [0..15]$), miniblock-half (1 = trial 1-7, 0 = trial 8-15) and the interaction term absolute score difference*miniblock-half. There was a significant main effect of absolute score difference (β = 0.14, SE = 0.008, p < 0.001) and miniblock-half (β = 0.29, SE = 0.039, p < 0.001). Importantly, the analysis revealed a significant interaction between miniblock-half and absolute score difference (β = -0.2, SE = 0.013, p < 0.001). This means that goal choice was more affected by the absolute score difference in the second half the miniblock compared to the first half. The analysis supports our conclusion that participants relied on a heuristic strategy preference when far from the goal.

2.6 Discussion

In the current study, we investigated how humans change the way they decide what goal to pursue while approaching two potential goals. To emulate real life temporally extended decision making scenarios of goal pursuit, we used a novel sequential decision making task. In this task environment, decisions of participants had deterministic consequences, but the options given to participants on each of the 15 trials were stochastic. This meant that especially during the first few trials, participants could not predict with certainty what goal was achievable. Using model-based analysis of behavioural data we find that most participants, during the initial trials, relied on computationally inexpensive heuristics and switched to forward planning only when closer to the final trial.

We inferred the transition from a heuristic action selection to action selection based on forward planning using a model parameter that captured participants' preference for pursuing both goals either in a sequential or parallel manner. This strategy preference had its strongest impact for the first few trials, when participants, due to the stochasticity of future offers, could not predict well which of the two available actions in a mixed trial would enable them to maximize their gain. This can be seen from Eq. 2.11 where two terms contribute to making a decision: the term containing the differential expected value (DEV) and the strategy preference θ. In our computational model, the DEV is the difference between the expected value of a sequential strategy choice and a parallel strategy choice. The DEV enables the agent to choose actions which maximize the average reward gain in a miniblock (see methods). Critically, this DEV is typically close to 0 in the first few trials, i.e. there is high uncertainty on what action is the best one. In this situation, the strategy preference mostly determines the action selection of the agent. In our model, we computed the DEV by using forward planning, where the agent hypothetically runs simulations through all remaining future trials until the end of a miniblock, i.e. to the 15th trial. The number of state space trajectories to be considered in these simulations scales exponentially with the number of remaining trials – and so does in principle the computational costs needed to simulate these trajectories. Therefore, full forward planning would be both prohibitively costly and potentially useless when the deadline is far away, rendering simpler heuristics (Soltani et al., 2016) the more appropriate alternative.

It is an open question what heuristic participants actually used. In our model, the strategy preference parameter simply quantifies a preference for a parallel or sequential strategy and biases a participant's action selection accordingly. This may mean that participants had a prior expectation whether they are going to reach G2 or just G1. Given this prior, participants could choose their action without any forward planning. In other words, to select an action in a mixed trial, participants simply assumed that they are going to reach, for example, G2. This simplifies action selection tremendously because, under the assumption that G2 will be reached, the optimal action is to use the parallel strategy at all times. To an outside observer, a participant with a strong preference for a parallel strategy may be described as overly optimistic, as this participant would choose g2-choices even if reaching G2 is not very likely, e.g. in the hard condition. Conversely, a participant with a strong preference for a sequential strategy may be described as too cautious, e.g. because that participant chooses one-goal actions in the easy condition (see S2.12 Fig for two example participants). Importantly, the difference in total reward between the agent and the participants is only about 5% (see Fig 2.3, E). This means that even though participants used a potentially suboptimal strategy preference, the impact on total reward is not that large. This is because, as we have shown, later in the miniblock, when $DEVs$ become larger and are more predictive of what goal can be reached, participants choose their actions accordingly. Although we do not quantify the relative costs of full forward planning versus the observed mixture of heuristic and forward planning, we assume that an average loss of 5% of the earnings is small as compared to the reduction of computational costs when using heuristics.

There were two important features of our sequential decision making task: The first was that we used a rather long series of 15 trials to model multiple goal pursuit, where typically sequential decision making tasks would use fewer trials, e.g. 2 in the two-step task (Daw et al., 2011) with common values around 5 (Korn & Bach, 2018) to 8 trials (Kolling et al., 2014; Schwartenbeck et al., 2015) per miniblock. The reason why we chose a rather large number of trials is that this effectively precluded the possibility that participants can plan forward and ensure that participants were exposed at least to some initial trials where they had to rely on other information than forward planning. This initial period when participants have to select actions without an accurate estimate of the future consequences of these actions is potentially most interesting for studying meta-decisions about how we use heuristics when detailed information about goal reaching probabilities is scarce. It is probably in this period of uncertainty during goal reaching, when internal beliefs and preferences have their strongest influence.

The second important feature of our task was that participants had to prioritize between two goals. This is a departure from most sequential decision making tasks, where there is typically a single goal, e.g. to collect a minimum number of points, where the alternative is a fail (Kolling et al., 2014). In our task, participants could reach one of two goals, which enables addressing questions about how participants select and pursue a specific goal, see also (Ballard et al., 2016). Our findings complement work investigating behavioural strategies for pursuing multiple goals, e.g. (Orehek & Vazeou-Nieuwenhuis, 2013), showing that pursuit strategies depend on environmental characteristics, subjective preferences and changes in context when getting closer to the goal. In line with our findings, a recent study (Juechems et al., 2019) showed that decisions whether to redress the imbalance between two assets or to focus on a distinct asset during sequential goal pursuit were best fit by a dynamic programming model with a limited time horizon of 7.5 trials (20 trials would be the optimum). In future research, the pursuit of multiple goals in sequential decision making tasks may also be a basis for addressing questions about cognitive control during goal-reaching, e.g. how participants regulate the balance between stable maintenance and flexible updating of goal representations (Goschke, 2014).

In the current experiment, time (trial within miniblock) was correlated with both, planning complexity (exponential growth of the planning tree) and the magnitude of $DEVs$ (S2.5 Fig). However, complexity and time can, in principle be dissociated. For example, a temporally distant goal might have only low planning complexity because one must consider only a few decision sequences leading to the goal. Conversely, a temporally proximate goal might have high planning complexity because of a large number of potential actions sequences that may lead to the goal. Moreover, in contrast to the current task, there might be situations, in which the early decisions matter most. This would be reflected in large $DEVs$ at beginning of the goal reaching sequence. In future research, by testing sequential tasks with varying transition structure, one could selectively test how DEV-magnitude, time and complexity influences the arbitration of forward planning and the use of heuristics.

It is unclear what mechanism made participants actually use a strategy preference different from zero in our task. It is tempting to assume that participants might have used their usual approach, which they might apply in similar real-life situations, to select their goal strategies when the computational costs of forward planning are high and the prediction accuracy is low. In other words, participants who had a preference for a parallel strategy might either show a tendency towards working on multiple goals at the same time or entertain the belief that tasks should be approached with an optimistic stance. Conversely, participants with a preference for a sequential strategy might have made good experiences with using a more cautious approach and would tend to pursue one goal after the other.

We would like to note that the proposed model does not explicitly model the arbitration between forward planning and heuristic decision making. The computational model to fit participant behaviour uses at its core full forward planning as the optimal agent does. The effect of strategy preference just changes the action selection result, but the underlying computation to determine the DEV is still based on forward planning. Clearly, if a real agent used our model, this agent would not save any computations because forward planning is still used for all trials. The open question is how an agent makes a meta-decision to not use goal-directed forward planning but to rely on heuristics and other cost-efficient action selection procedures (Boureau et al., 2015). To make this meta-decision, an agent cannot rely on the DEV because this value is computed by forward planning. An alternative way would be to use an agent's prior experience to decide that the goal is still too temporally distant to make an informed decision with an acceptable computational cost. Such a meta-decision would depend on several factors, e.g. the relevance of reaching G2, intrinsic capability and motivation of planning forward, or a temporal distance parameter which signals urgency to start planning forward. In the future we plan to develop such meta-decision-making models and predict the moment at which forward planning takes over the action selection process.

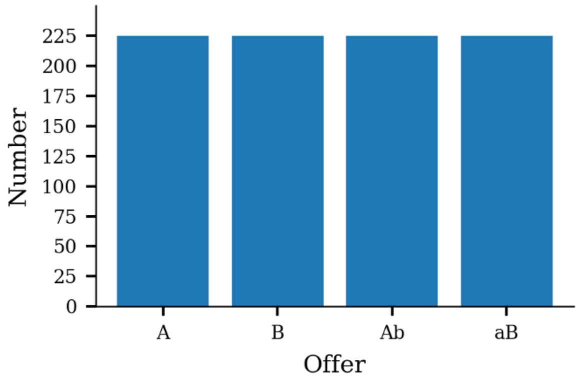
It is also possible that participants use, apart from simple heuristics, other approximate planning strategies to reduce computational costs. For example, one could sample only a subset of sequences to compute value estimates. Indeed, in another study it was found that participants prune a part of the decision tree in response to potential losses, even if this pruning was suboptimal (Huys et al., 2012). Another important point is that the planning process itself might be error-prone and therefore value calculations over longer temporal horizons may be noisier. This could presumably account for temporal modulations of the precision parameter β. In future work one could test for evidence of alternative planning algorithms that allow to sample subsets of (noisy) forward planning trajectories to further delineate how humans deal with computational complexity in goal-directed decision scenarios. Furthermore, in the current analysis, we tested a model, where we fitted parameters separately for early, middle and late trials of the miniblock, but found that this time-variant model had lower model evidence compared to the constant model. We interpreted these results as further evidence that the described constant parameterization is sufficient to describe a hidden shift from using a heuristics to forward planning. Nevertheless, it is possible that there is an alternative dynamic model that explains the data better. In our analysis we decided to split one miniblock arbitrarily into three time bins - one could have also considered other splits, e.g. into two. Furthermore, it is possible that these time bins have a different structure from subject to subject (e.g. for some subject, the first ten trials will be associated with one parameter value and the last 5 with another and for some other subject it could be the first 7 and last 8 trials). However, we will leave the detailed exploration of dynamic model parametrisation for the future.

Taken together, the present research shows that over prolonged goal-reaching periods, individuals tend to behave in a way that approaches the behaviour of an optimal agent, with noticeable differences early in the goal-reaching period, but nearly optimal behaviour when the goal is close. It also highlights the potential of computational modelling to infer the decision parameters individuals use during different stages of sequential decision making. Such models may be a promising means to further elucidate the dynamics of decision making in the pursuit of both laboratory and everyday life goals.

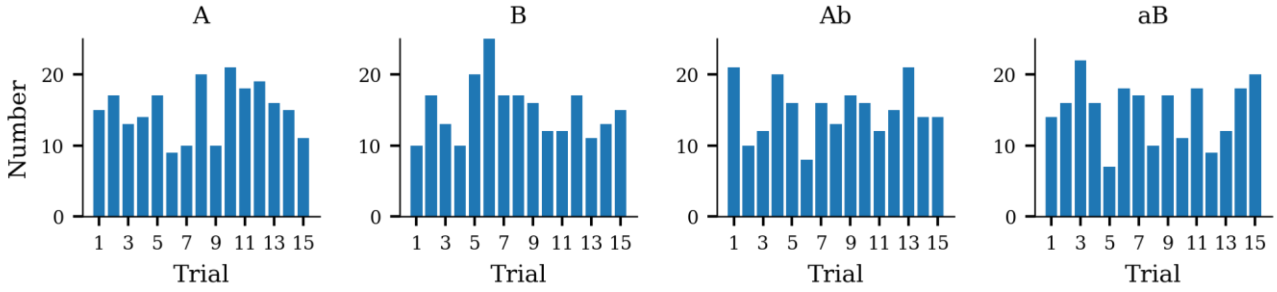
2.7 Supporting information

S2.1 Table. Classification of accept-wait responses into either two-goal-choices (g2) or one-goal-choices (g1).

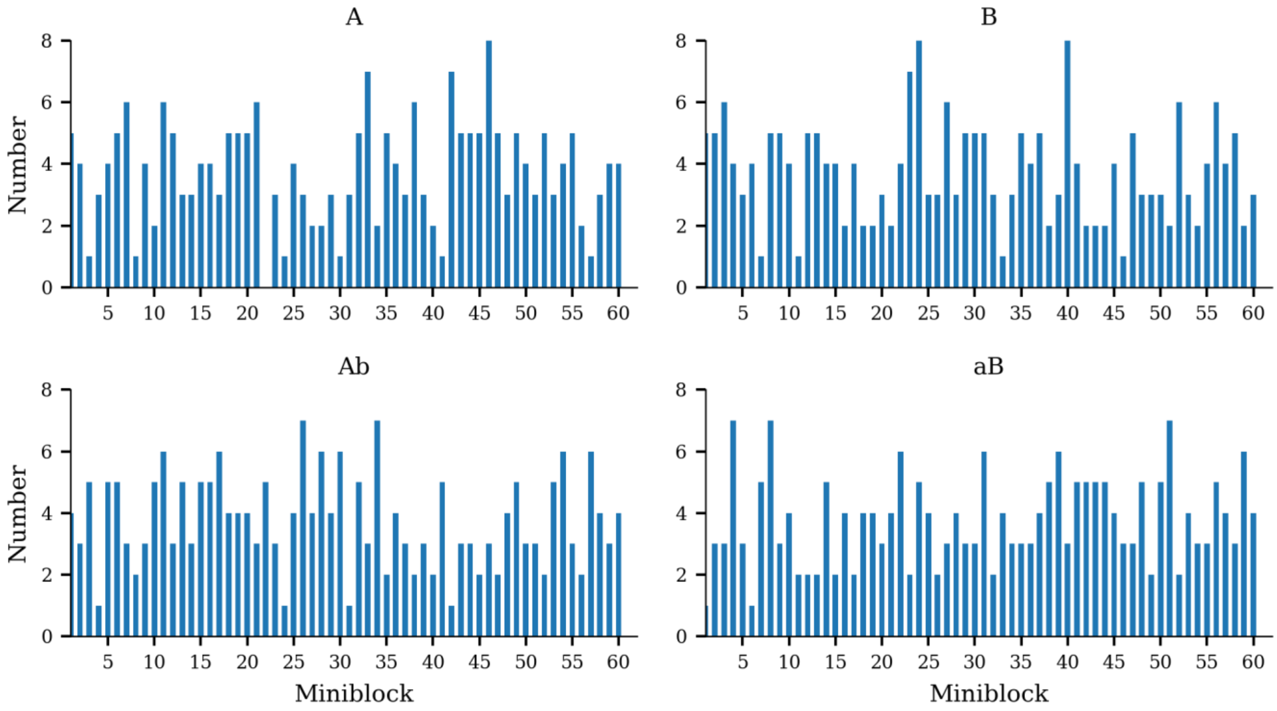
Offer	Points	Response	Classification
Ab	$Pts_t^A - Pts_t^B > 1$	accept	g1
Ab	$Pts_t^A - Pts_t^B > 1$	wait	g2
Ab	$Pts_t^A - Pts_t^B < -1$	accept	g2
Ab	$Pts_t^A - Pts_t^B < -1$	wait	g1
Ab	$Pts_t^A - Pts_t^B = 1$	accept	g1
Ab	$Pts_t^A - Pts_t^B = 1$	wait	g2
Ab	$Pts_t^A - Pts_t^B = -1$	accept	nan
Ab	$Pts_t^A - Pts_t^B = -1$	wait	nan
Ab	$Pts_t^A - Pts_t^B = 0$	accept	g1
Ab	$Pts_t^A - Pts_t^B = 0$	wait	g2
aB	$Pts_t^A - Pts_t^B > 1$	accept	g2
aB	$Pts_t^A - Pts_t^B > 1$	wait	g1
aB	$Pts_t^A - Pts_t^B < -1$	accept	g1
aB	$Pts_t^A - Pts_t^B < -1$	wait	g2
aB	$Pts_t^A - Pts_t^B = 1$	accept	nan
aB	$Pts_t^A - Pts_t^B = 1$	wait	nan
aB	$Pts_t^A - Pts_t^B = -1$	accept	g1
aB	$Pts_t^A - Pts_t^B = -1$	wait	g2
aB	$Pts_t^A - Pts_t^B = 0$	accept	g1
aB	$Pts_t^A - Pts_t^B = 0$	wait	g2



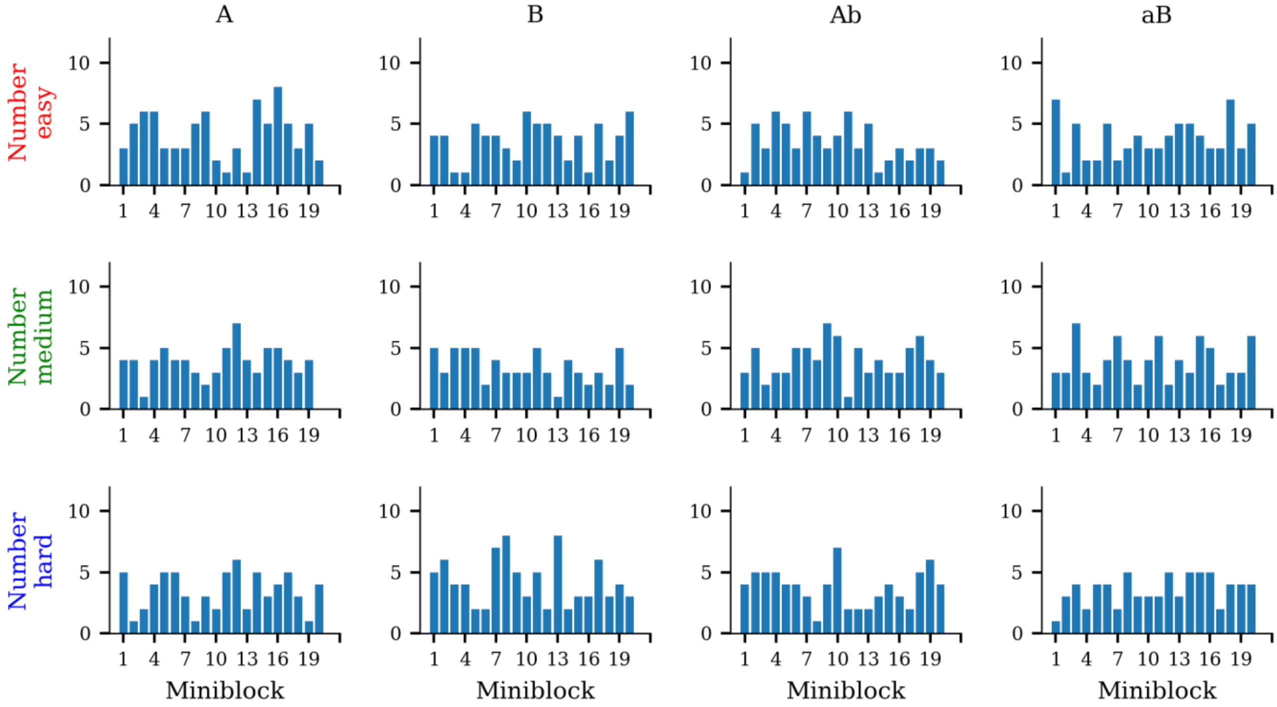
S2.1 Fig. Occurrence of offer types across all 900 trials.



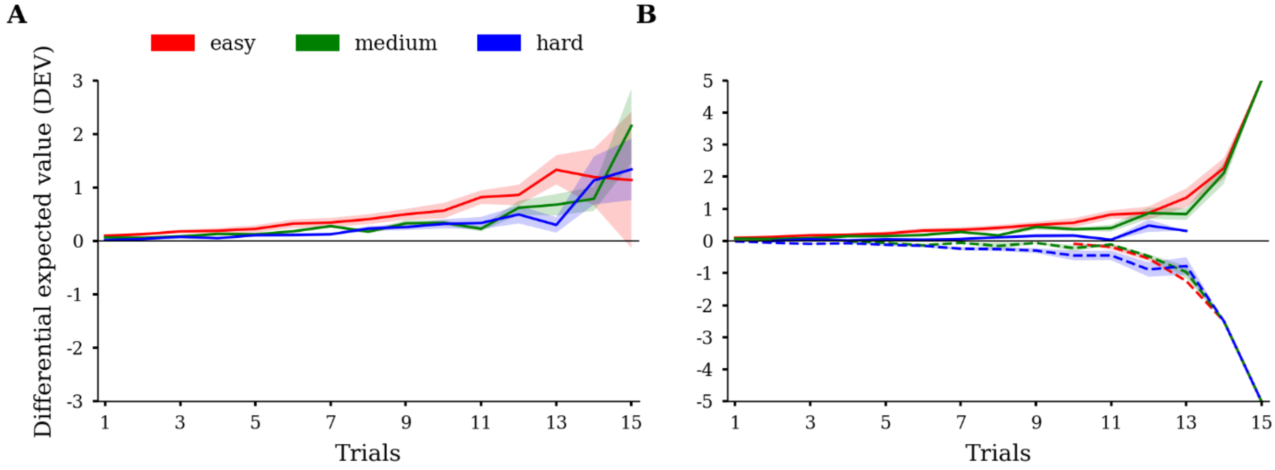
S2.2 Fig. Occurrence of offer types binned with respect to trial.



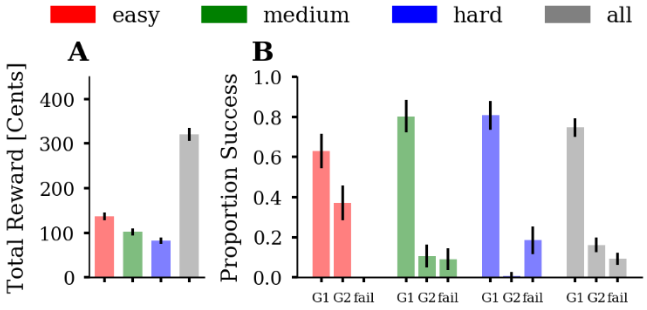
S2.3 Fig. Occurrence of offer types binned with respect to miniblock.



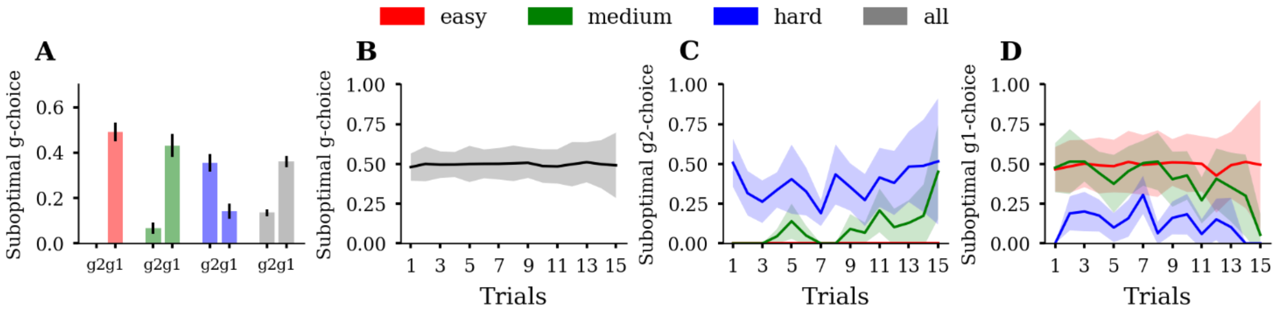
S2.4 Fig. Occurrence of offer types binned with respect to miniblock and difficulty.



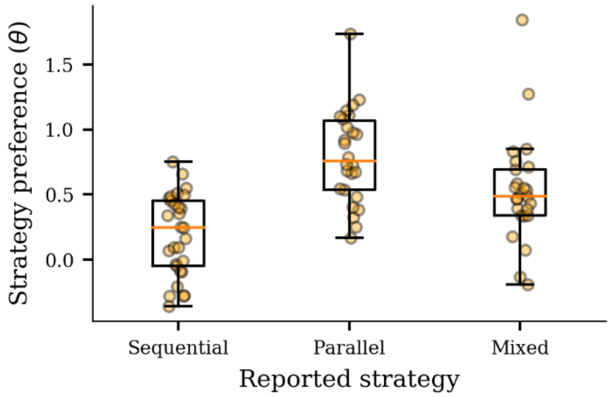
S2.5 Fig. Average absolute (A) and signed (B) differential expected value (DEV) per trial and condition. Discount and reward ratio had been fixed (γ = 1, κ = 1). Average absolute $DEVs$ at the beginning of the miniblock are smaller than in the end, indicating the relative importance of decisions close to the final trial of miniblocks. Conditions are colour coded. The shaded areas represent SD.



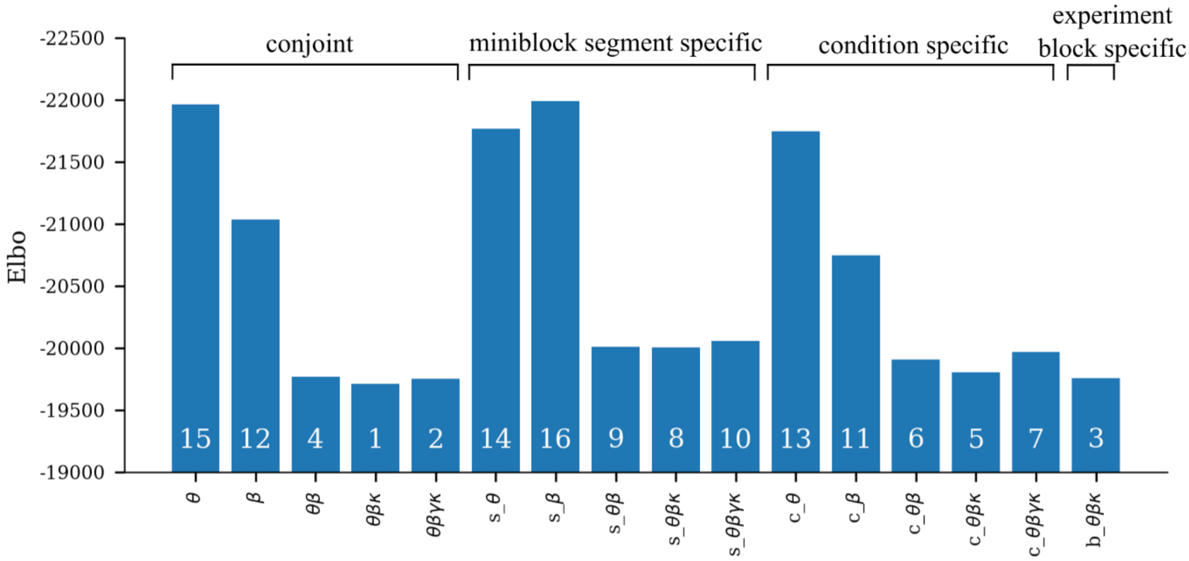
S2.6 Fig. Simulated goal success and total reward of a random agent that always accepts basic offers but guesses for mixed offers ($\theta = 0$, $\beta \to 0$, $\gamma = 1$, $\kappa = 1$). (A) Average total reward across agent instances (n =1000). **(B)** Proportion of successful goal-reaching, averaged across agent instances, for each of the three conditions. We plot the proportion of reaching, at the end of a miniblock, a single goal (G1), both goals (G2), or no goal (fail). The random agent achieves fewer G2-successes in easy and medium than the participants but fails more often in medium and hard. The three conditions are colour-coded (easy = red, medium = green, blue = hard) and the average over conditions is shown in grey. Error bars depict SD.



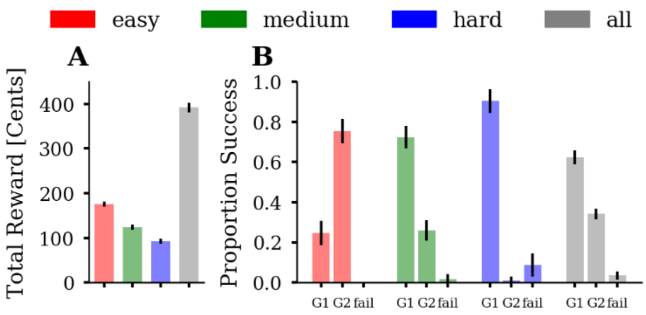
S2.7 Fig. Simulated suboptimal g-choices of a random agent that always accepts basic offers but guesses for mixed offers ($\theta = 0, \beta \to 0, \gamma = 1, \kappa = 1$). (A) Proportions of suboptimal g1-choices (g1) and suboptimal g2-choices (g2), averaged over agent instances (n =1000). The random agent makes many suboptimal g1-choices in the easy and medium and many suboptimal g2-choices in the hard conditions. Summing together g1 and g2 yields approximately 50% suboptimal g-choices. **(B)** Suboptimal g-choices as a function of trial averaged over agent instances. The random agent makes approximately 50% suboptimal g-choices across all trials in the miniblock. If participants use non-random response strategies, i.e. planning or heuristics, their pattern of suboptimality across trials should deviate from the straight-line pattern of the random agent. **(C)** Suboptimal g2-choices as a function of trial averaged over agent instances. **(D)** Suboptimal g1-choices as a function of trial averaged over agent instances. Summing together g1 (D) and g2 (C) yields approximately 50% suboptimal g-choices across trials. Error bars and shaded areas depict SD. Conditions are colour coded.



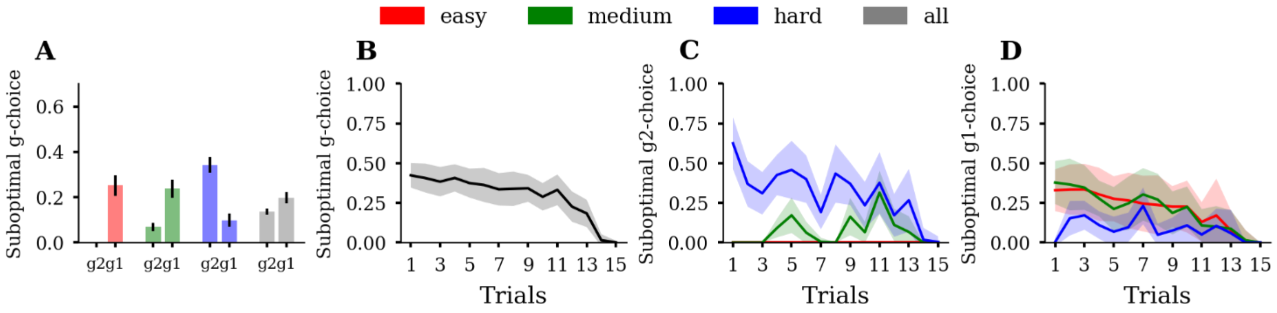
S2.8 Fig. Qualitative comparison of participants' reported strategy use and fitted strategy preference parameter. Participants who reported the use of a sequential strategy had lower estimated strategy preference, including the most negative values, than participants who reported the use of a parallel strategy. Participants who reported mixed use of a parallel and sequential strategy had greater strategy preference than the sequential group but lower estimates than the parallel group. The plot shows 80 of 89 participants whose verbal reports matched with one of the three strategy categories.



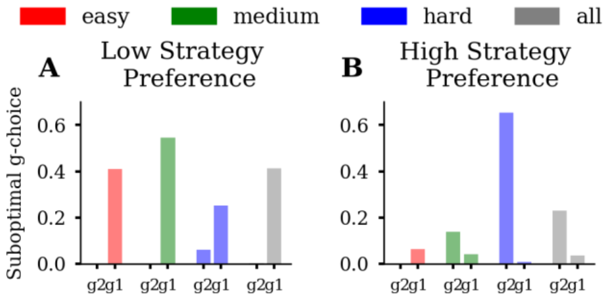
S2.9 Fig. Comparing Elbo (evidence lower bound) between different model variants. White numbers represent the rank from highest to lowest Elbo. Model comparisons showed that the three parameter model (θ, β, κ) had the highest model evidence. Adding γ did not increase model evidence $(elbo_{\theta\beta\kappa} - elbo_{\theta\beta\gamma\kappa} = -44)$. Estimating model parameters separately for miniblock segments (trial 1-5, trial 6-10, trial 11-15; prefix 's_' in the figure) had lower model evidence compared to the winning model $(elbo_{\theta\beta\kappa} - elbo_{s\_\theta\beta\kappa} = -294)$. Estimating model parameters separately for conditions (easy, medium, hard; prefix 'c' in the figure) had lower model evidence compared to the winning model $(elbo_{\theta\beta\kappa} - elbo_{c\_\theta\beta\kappa} = -94)$. Estimating model parameters separately for experiment blocks (miniblock 1-20, miniblock 21-40, miniblock 41-60; prefix 'b' in the figure) had also lower model evidence compared to the winning model $(elbo_{\theta\beta\kappa} - elbo_{s\_\theta\beta\kappa} = -48)$. Bars in the plot depict Elbo averaged over the last 20 posterior samples.



S2.10 Fig. Posterior predictive checks: Simulated goal success and total reward closely resemble observed participant behaviour. **(A)** Average total reward across samples (n = 1,000). **(B)** Proportion of successful goal-reaching, averaged across samples, for each of the three conditions. We plot the proportion of reaching, at the end of a miniblock, a single goal (G1), both goals (G2), or no goal (fail). The three conditions are colour-coded (easy = red, medium = green, blue = hard) and the average over conditions is shown in grey. Error bars depict SD. Data were generated using 1,000 posterior samples from the group hyper parameters.



S2.11 Fig. Posterior predictive checks: Simulated suboptimal g-choices closely resemble observed participant behaviour. (A) Proportions of suboptimal g1-choices (g1) and suboptimal g2-choices (g2), averaged over samples (n =1,000). **(B)** Suboptimal g-choices as a function of trial averaged over samples. **(C)** Suboptimal g2-choices as a function of trial averaged over samples. **(D)** Suboptimal g1-choices as a function of trial averaged over samples. Error bars and shaded areas depict SD. Conditions are colour coded. Data were generated using 1,000 posterior samples from the group hyper parameters.



S2.12 Fig. Comparison of suboptimal g-choices between a low strategy preference and high strategy preference participant. The plot shows proportions of suboptimal g1-choices (g1) and suboptimal g2-choices (g2) **(A)** of the participant with the lowest fitted strategy preference ($\theta = -0.36$) and **(B)** of the participant with the highest fitted strategy preference ($\theta = 1.84$). The low strategy preference participant prefers a sequential strategy leading to suboptimal g1-choices in the easy and medium condition. The participant with a high strategy preference parameter prefers a parallel strategy, resulting in a few suboptimal g1-choices in easy in and medium but a large number of suboptimal g2-choices in the hard condition.

S2.1 Text: Task instructions (translated from German)

- Dear participant, your task in this experiment is to reach goals. Within a block, consisting of 15 trials, you can either reach goal A, goal B or both goals at the same time. For one reached goal you will gain additional 5 Cents and for two reached goals additional 10 Cents. Your task is to obtain as much money as possible.
- To reach goals, you must collect points. You can get points by accepting an offer. Some offers however, might have a negative effect on the state of a goal. Your task is to decide in every trial, whether to accept an offer or wait for the next offer. Press "up arrow" to accept an offer and "down arrow" to wait.
- Important: Please decide deliberately but speedily. If you decide too slowly, you will get a notification. After every 5 notifications, 50 Cents will be subtracted from your bonus-payout. (The experiment starts with a training phase, in which no money can be lost.)
- More about the goals: Your goal progress will be represented by a bar, which is labelled with A or B. A goal counts as achieved, if one of the bars reaches or surpasses the white horizontal mark. The goal state will be evaluated after the end of the 15 trials.
- More about the offers: There are 4 different offers – A, B, Ab an aB. All offers have the same occurrence probability of 25%. The offers differ with respect to their effect on the goal state. A increases the A-bar by one point. B increases the B-bar by one point. Ab increases the A-bar by one point and subtracts one point from the B-bar. aB increases the B-bar by one point and subtracts 1 point from the A-bar.
- Initial conditions: At the beginning of the block, you already have some A- and B-points. The amount of initial points varies from block to block.

S2.1 Movie. Simulated goal success and total reward where the precision parameter β varies between 0.25 and 3 with θ, γ, and κ sampled from their fitted population mean. (A) Average total reward across agent instances (n =1,000). An increase in β increases total reward obtained in the easy and medium but decreases total reward in the hard condition. **(B)** Proportion of successful goal-reaching, averaged across agent instances, for each of the three conditions. We plot the proportion of reaching, at the end of a miniblock, a single goal (G1), both goals (G2), or no goal (fail). An increase in β increases G2 success rate in easy and medium but also increases fail rate in medium and hard. The three conditions are colour-coded (easy = red, medium = green, blue = hard) and the average over conditions is shown in grey. Error bars depict SD. See online material.

S2.2 Movie. Simulated suboptimal g-choices where the precision parameter β varies between 0.25 and 3 with θ, γ, and κ sampled from their fitted population mean. (A) Proportions of suboptimal g1-choices (g1) and suboptimal g2-choices (g2), averaged over agent instances (n =1000). An increase in β decreases suboptimal g1- and g2-choices. **(B)** Suboptimal g-choices as a function of trial averaged over agent instances. The influence of β and the associated decrease of suboptimal g-choices successively increases towards the end of the miniblock. Suboptimal g-choices in the first half of the miniblock are largely unaffected by the β parameter. **(C)** Suboptimal g2-choices as a function of trial averaged over agent instances. An increase in β decreases suboptimal g2-choices late in the miniblock in medium and hard but not in easy. **(D)** Suboptimal g1-choices as a function of trial averaged over agent instances. An increase in β decreases suboptimal g1-choices late in the miniblock in easy and medium but not in hard. Error bars and shaded areas depict SD. Conditions are colour coded. See online material.

S2.3 Movie. Simulated goal success and total reward where the strategy preference parameter θ varies between -1 and 1 with β, γ, and κ sampled from their fitted population mean. (A) Average total reward across agent instances (n =1000). An increase in θ increases total reward obtained in easy and medium but decreases total reward in hard. **(B)** Proportion of successful goal-reaching, averaged across agent instances, for each of the three conditions. We plot the proportion of reaching, at the end of a miniblock, a single goal (G1), both goals (G2), or no goal (fail). An increase in θ increases G2 success rate in easy and medium but also increases fail rate in medium and hard. The three conditions are colour-coded (easy = red, medium = green, blue = hard) and the average over conditions is shown in grey. Error bars depict SD. See online material.

S2.4 Movie. Simulated suboptimal g-choices where the strategy preference parameter θ varies between -1 and 1 with β, γ, and κ sampled from their fitted population mean. (A) Proportions of suboptimal g1-choices (g1) and suboptimal g2-choices (g2), averaged over agent instances (n =1000). An increase in θ decreases suboptimal g1- choices and increases suboptimal g2-choices. Suboptimal g1-choices decrease more in easy and medium than in hard. Suboptimal g2-choices decrease more in hard than in easy and medium. **(B)** Suboptimal g-choices as a function of trial averaged over agent instances. A change in θ affects the number of suboptimal g-choices made at the beginning but not at the end of the miniblock. For $\theta > 0$ suboptimal g-choices further decrease, because g2-choices are often optimal in easy and medium. **(C)** Suboptimal g2-choices as a function of trial averaged over agent instances. An increase in θ increases suboptimal g2-choices early in the miniblock, predominantly in the hard condition. **(D)** Suboptimal g1-choices as a function of trial averaged over agent instances. An increase in θ decreases suboptimal g1-choices early in the miniblock, predominately in easy and medium. Error bars and shaded areas depict SD. Conditions are colour coded. See online material.

S2.1 Notebook. Parameter recovery simulations. See online material.

3 Study 2: Forward planning driven by context dependent conflict processing in anterior cingulate cortex

3.1 Abstract

Forward planning is often essential to achieve goals over extended time periods. However, forward planning is typically computationally costly for the brain and should only be employed when necessary. The explicit calculation of how necessary forward planning will be, is in itself computationally costly. We therefore assumed that the brain generates a mapping from a particular situation to a proxy of planning value to make fast decisions about whether to use forward planning, or not. Moreover, since the state space of real world decision problems can be large, we hypothesized that such a mapping will rely on mechanisms that generalize sets of situations based on shared demand for planning. We tested this hypothesis in an fMRI study using a novel complex sequential task. Our results indicate that participants abstracted from the set of task features to more generalized control contexts that govern the balancing between forward planning and a simple response strategy. Strikingly, we found that correlations of conflict with response time and with activity in the dACC were dependent on context. This context dependency might reflect that the cognitive control system draws on category-based cognition, harnessing regularities in control demand across task space to generate control contexts that help reduce the complexity of control allocation decisions.

3.2 Introduction

Many decisions have far-reaching consequences for the future, as they affect both internal bodily and external environmental states, in turn often conditioning potential future actions. Therefore, to achieve any long-term goals, people have to consider the future in some way. This can be achieved by planning multiple steps into the future to estimate the effects of potential action sequences (K. J. Miller & Venditto, 2020; D. A. Simon & Daw, 2011; Tolman, 1948). However forward planning comes at a cost of using time and cognitive capacities, therefore people should only plan ahead when the benefits outweigh the costs and rely on fast and frugal strategies otherwise (Gershman et al., 2015; Gigerenzer & Gaissmaier, 2011; Kool et al., 2017; Lieder & Griffiths, 2020; Shenhav et al., 2013; Shenhav et al., 2017).

An intriguing question is how the brain controls when to engage in forward planning and when to use simpler strategies. Planning is often seen as one of the core functions of cognitive control (M. M. Botvinick & Cohen, 2014; Goschke, 2013; E. K. Miller & Cohen, 2001) and therefore the neural mechanisms involved in the regulation of cognitive control might be similarly involved in the regulation of planning. A classic hypothesis proposes that the dACC plays a central role in cognitive control by monitoring processing conflicts that serve as a signal for the need for additional control (M. M. Botvinick et al., 2001). Empirical evidence supports the involvement of the dACC in conflict processing in response interference tasks (Kerns et al., 2004; E. H. Smith et al., 2019), value-based decision making (Pochon et al., 2008) and recently in tasks that require people to plan multiple steps into the future (Economides et al., 2015; Korn & Bach, 2018; Schwartenbeck et al., 2015).

In multi-step tasks, however, an intricate computational problem becomes apparent. How can the brain control the use of forward planning in a way that maximizes long-term benefits, without having to compute these benefits by forward planning beforehand? One solution to this paradox might be for people to generate a mapping from a particular situation to a proxy of the value of planning that allows them to quickly access the planning values later (D. G. Lee & Daunizeau, 2021; Lieder et al., 2018). Moreover, because the state space for real-world decision problems can be very large, it is unlikely that people learn a value for every possible combination of states. Rather, they might use certain task features to generalize clusters of states into particular contexts for which values are learned (Lieder et al., 2018).

Here, we tested this principle in an fMRI study using a novel sequential decision making task. In the task participants had to plan ahead to earn points by accepting offers while managing a limited energy budget. Importantly, we designed the task such that situations with different levels of the demand for planning occurred. With 448 possible combinations of task features and four different offers participants could choose from, our task was quite complex. We therefore assumed that participants used a simplified representation of planning value during control allocation decisions. An initial analysis of choice frequencies showed that participants used a repetitive choice pattern for two of the options, while responses were more balanced for the two other offers. From these choice patterns, we hypothesized that participants generated two different groups of offer-dependent representations of planning value. We refer to these two groups as control contexts (or context for short), with one context coding for a high a priori need for planning and the other context coding for a low a priori need for planning. To further test the control context hypothesis, we analysed response times and fMRI data using a specific conflict measure as a proxy for the value of forward planning. We found that correlations of conflict with response time and with BOLD-activity in the dACC were dependent on the context. Our results provide initial evidence for a mechanism by which the brain harnesses regularities in the value of planning across tasks space to construct control contexts that facilitate efficient allocation of control in complex tasks. Future research should further develop and confirm these initial findings by testing formal models of arbitration which incorporate structured representations of planning value.

3.3 Methods

3.3.1 Participants

Forty participants took part in the experiment (22 women, mean age = 24.4, SD = 4.6). Reimbursement was a fixed amount of 14€ or class credit plus a performance-dependent bonus (mean bonus = 6.62€, SD = 0.39). The bonus was calculated as a linear function of the accumulated points in the experiment. The study was approved by the Institutional Review Board of the Technische Universität Dresden and conducted in accordance to ethical standards of the Declaration of Helsinki. All participants were informed about the purpose and the procedure of the study and gave written informed consent prior to the experiment. All participants had normal or corrected-to-normal vision.

3.3.2 Data availability

Data and analysis code used in this article is publicly available at https://doi.org/10.5281/zenodo.5112965. The repository includes: raw behavioural data and fMRI statistical maps underlying Figures 3.4 and 3.5; source code for reproducing Figures 3.2, 3.3, S3.2, and S3.3; source code implementing the model fitting, validation and comparison procedures.

3.3.3 Experimental task

To investigate the context dependency of people's propensity to plan ahead, we designed a novel open-ended sequential decision task in which participants had to accumulate as many points as possible. We induced a necessity for planning by introducing a limited but replenishable energy resource, which was required to accept offers. Planning was further encouraged by introducing variation around the energy cost of accepting and by explicitly informing participants about how these costs would change in the future. Unlike in experiments with a fixed deadline (Economides et al., 2014; Ott et al., 2020; Schwartenbeck et al., 2015), the open-ended nature of the task provided for every trial an opportunity to plan ahead multiple steps into the future. Importantly, the task featured both situations in which planning was crucial to decide between the accept and reject option, as well as situations that could be sufficiently solved by a simple heuristic.

In detail, the temporal structure of the task comprised three levels, the single trial, the current segment, and the segment pair of the current and next segment. One segment consisted of four trials. In a single trial, participants could either accept or reject an offer (selected by either a left or right button press), where accepting the offer increased points by an indicated amount but decreases energy by one or two units, depending on condition, see below. Rejecting always increased energy by one. There were four equally probable offers, displayed as one, two, three or four trophies in the middle of the screen. Accepting an offer increased points by the respective number of trophies, thereby advancing the yellow point bar at the top of the screen (Fig Fig **3.1**, A). The energy costs of accepting varied between one, in low-cost segments (LC), and two in high-cost segments (HC). The energy budget with a minimum of zero and a maximum of six was displayed as a blue bar at the bottom of the screen. Initial energy in the first trial of the experiment was three. If participants had maximum energy and chose to reject, no further energy was added, and the next trial started. If participants accepted an offer with too little energy, no points were awarded, a warning was displayed, and the next trial started. Participants were informed about the energy cost of the current and the future segment by two symbols in the bottom right corner. The left symbol informed about both the energy cost of the current segment and the current trial number in the segment. The right symbol informed about the energy cost of the future segment. One flash indicated a low-cost segment and two flashes a high-cost segment (Fig 3.1, A).

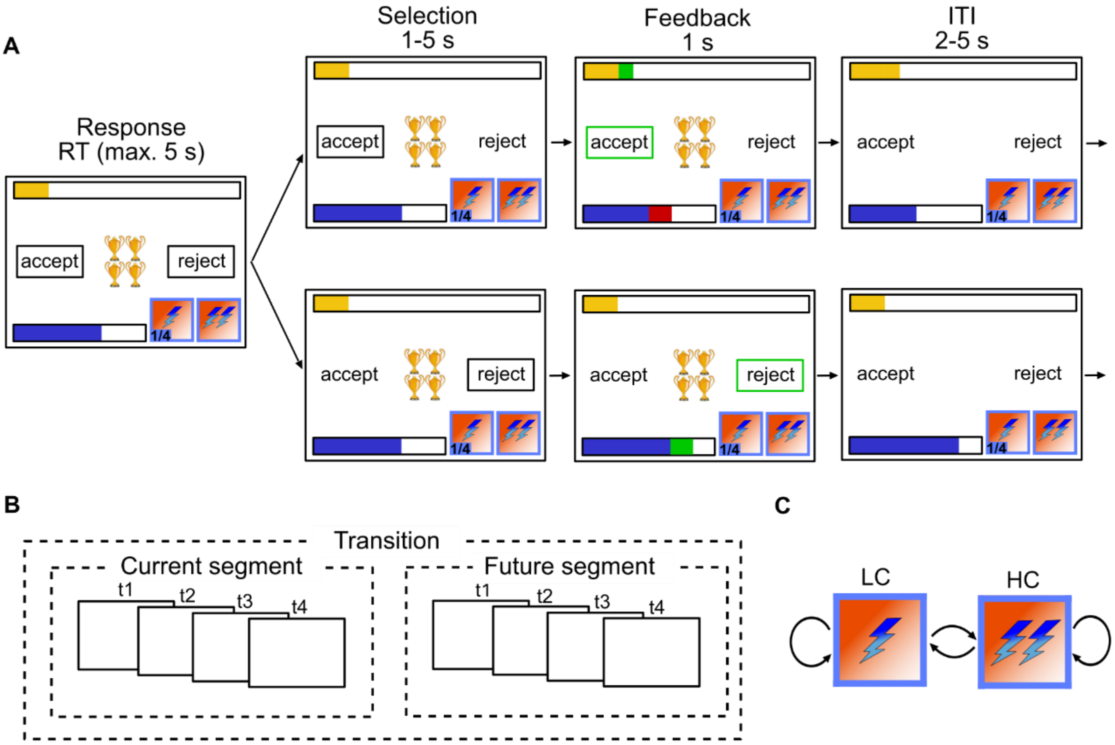


Fig 3.1. Experimental Task. (A) Timeline of a single trial. Participants could accept an offer ($O = \{1,2,3,4\}$) displayed in the middle of the screen to collect points (top yellow bar). Accepting an offer was associated with an energy cost. The current energy cost within a 4-trial-segment was indicated by the left symbol on the bottom right of the screen. One flash indicates an energy cost of 1 and two flashes an energy cost of 2. The right symbol on the bottom right indicates the energy cost of the next segment. Participants could choose the reject option to replenish the energy budget (bottom

blue bar) by 1. **(B)** Temporal structure of the task: Four single trials formed a segment. Two consecutive segments formed a segment pair. **(C)** A segment could feature one of two different energy costs, low energy costs (LC) or high energy costs (HC). There were four different segment transitions LC/LC, LC/HC, HC/LC and HC/HC.

The experiment included a training session outside the MRI scanner (for task instructions see S3.1 Text) and three sessions inside. The training session comprised 144 trials across 36 segments with nine repetitions for each of the four possible transitions (Fig 3.1, C). The fMRI experiment comprised 240 trials across 60 segments with 15 repetitions per segment transition. On average, participants took 40 minutes to complete the fMRI experiment. The fMRI experiment was split into three sessions between which participants rested for two to three minutes without leaving the scanner. The sequence of segments and offers was pseudorandomized and identical for all participants. Segment sequences were generated such that each of the four segment transitions (Fig 3.1, C) was sampled equally often. Similarly, offer sequences were generated such that the frequency of offers was balanced within segment transitions (raw behavioural data with all details about the offer sequences can be found at https://doi.org/10.5281/zenodo.5112965).

The timing of stimulus events in the fMRI experiment was as follows (see also Fig. 3.1, A): each trial started with a fixation cross (0.5 seconds) in the middle of the screen to prepare participants for the upcoming decision. In the response phase, the offer appeared, and the choice options were surrounded by a frame to indicate that a decision is required. If participants did not respond within 5 seconds, they were timed out with a warning message, and the next trial began. In the selection phase (1-5 seconds, uniformly sampled and rounded to first decimal), the frame surrounding the unchosen option disappeared. In the feedback phase (1 second, fixed), energy or point changes were displayed, and the frame surrounding the chosen option turned green (or accordingly red if the energy budget was too low for accepting). In the intertrial interval (2-5 seconds, uniformly sampled and rounded to first decimal) choice options were unframed and the offer disappeared (Fig 3.1, A). The training version of the stimulus was identical to the fMRI version except that there was no timeout for the response phase, there was not a selection phase and the intertrial interval was fixed to 1 second.

3.3.4 Computational model of choice behaviour

We considered three different computational-cognitive models of how participants select their responses. Firstly, following one standard assumption about participants' behaviour in such sequential decision making tasks, participants may have used forward planning across the current and the next segment ('planning strategy') to estimate the expected value of either accepting or rejecting (e.g. Kolling et al., 2014; Schwartenbeck et al., 2015). Secondly, in contrast, participants may have used some sort of heuristic to reduce the number of computations involved. An obvious choice for such a heuristic is the simple strategy to just base the accept/reject decision on the offer value ('simple strategy'). For example, participants may simply always reject the lower offer values 1 and 2, and always accept offer values 3 and 4, provided there was enough energy left to accept the offer. Thirdly, we also considered a hybrid strategy of these two extremes, where participants may use a mixture between forward planning and simple strategy ('hybrid strategy'). Note that our modelling approach relies on logistic regression and does not describe per se a process of how the brain may balance between forward planning and a simple strategy. Rather, the computational approach enables us to test for evidence that participants rely on (i) forward planning, (ii) a simple strategy or (iii) a mixture between these two extremes.

*Planning strategy model (PM)***.** Clearly, if participants used a greedy strategy of accepting all offers in the first few trials of the task, they will quickly run out of energy and might not be able to accept better offers in future trials. Therefore, to maximize the accumulated points, one has to plan ahead, anticipating future actions, energy costs and reward opportunities. To implement such a planning strategy, we assumed a finite horizon until the end of the next future segment since participants were only explicitly informed about the energy costs of the current and the next future segment. As each segment had four trials each, this resulted in a horizon of maximally 8 trials and minimally 5 trials, i.e. when a participant has to select the decision for the fourth trial of the current segment. To derive a policy that maximizes expected reward over this horizon we formalised our task as a Markov Decision Process (MDP) (Puterman, 2014; Sutton & Barto, 2018). We define

$$MDP = (T, S, A, p_t(s'|s,a), r_t(s,a)) \tag{3.1}$$

where $T = \{1, 2, \dots, 8\}$ is the set of trials and $S = E \times O$ is the set of states with $E = \{0,1, \dots, 6\}$ the set of possible energy levels and $O = \{1,2,3,4\}$ the set of offer values. $A = \{accept, reject\}$ is the set of actions.

Since participants successfully completed a training session and received detailed task instructions (S3.1 Text) prior to the main experiment, we assume that participants understood the rules of the task. In the model, this knowledge is represented by the transition probability $p_t(s' = (e', o')|s = (e, o), a)$, which is the probability to transition to a new state s' given the current state s (consisting of the offer value o and the current energy e) and the selected action a. Probabilities for allowed state transitions satisfy

$$p_t(e' = e - cc, o'|e, o, a = accept) = p(o), \qquad e > cc, for\ all\ t \leq 4$$
$$p_t(e' = 0, o'|e, o, a = accept) = p(o), \qquad e < cc, for\ all\ t \leq 4$$

$$p_t(e' = e - fc, o'|e, o, a = accept) = p(o), \qquad e > fc, for\ all\ t > 4$$
$$p_t(e' = 0, o'|e, o, a = accept) = p(o), \qquad e < fc, for\ all\ t > 4$$
$$p_t(e' = e + 1, o'|e, o, a = reject) = p(o), \qquad e < 6, for\ all\ t$$
$$p_t(e' = e, o'|e, o, a = reject) = p(o), \qquad e = 6, for\ all\ t \tag{3.2}$$

where $p(o)$ is the discrete uniform distribution over the possible offer values, cc is the energy cost in the current segment (1 or 2) and fc is the energy cost in the future segment (1 or 2). We model the different segment transitions $LC \to LC, HC \to LC, LC \to HC\ and\ HC \to HC$ as separate MDPs, substituting the respective values for cc and fc.

Immediate rewards, corresponding to the offer value, are generated upon successful acceptance. Formally, the reward function satisfies

$$r_t(s = (e, o), a = accept) = o, \qquad if\ e \geq cc, for\ all\ t \leq 4$$
$$r_t(s = (e, o), a = accept) = 0, \qquad if\ e < cc, for\ all\ t \leq 4$$
$$r_t(s = (e, o), a = accept) = o, \qquad if\ e \geq fc, for\ all\ t > 4$$
$$r_t(s = (e, o), a = accept) = 0, \qquad if\ e < fc, for\ all\ t > 4$$
$$r_t(s = (e, o), a = reject) = 0, \qquad for\ all\ t \tag{3.3}$$

To determine the optimal policy that maximizes the expected reward over the current and future segment, the PM uses backward induction. The algorithm was implemented as follows :

1. Set $t = 9$ (if in the first trial of the segment) and define the state values after the decision in the final trial. To ensure that energy units left over after the eighth trial are considered to have utility, each remaining energy unit was multiplied by the quotient of the average offer value divided by the average energy costs.

$$V_{t=9}(s) = \frac{2.5}{1.5} e = 1.\overline{6}\, e \quad for\ all\ s_{t=9} \in S \tag{3.4}$$

2. Set $t = t - 1$ and compute state-action values

$$Q_t(s, a) = r_t(s, a) + \sum_{s' \in S} p_t(s'|s, a) V_{t+1}(s') \tag{3.5}$$

3. Update state values

$$V_t(s) = \underset{a}{\operatorname{argmax}}(Q_t(s, a)) \tag{3.6}$$

4. If t = 1 stop. Otherwise, continue with step 2

Action selection in the PM relies on a decision variable (DV) computed as the difference between the optimal state-action values

$$DV_{plan} = Q_t(s, a = accept) - Q_t(s, a = reject) \tag{3.7}$$

For a given state, positive values of DV indicate a greater long-term expected reward for accepting and negative values of DV indicate greater long-term expected reward for rejecting.

Using a logistic regression approach, we define the probability to accept as

$$p(accept) = \frac{1}{1 + e^{-\eta}} \tag{3.8}$$

where

$$\eta = \beta_{plan}\, DV_{plan} + \theta_{basic}\, I_{basic} + \theta_{maxE}\, I_{maxE} + \theta_{minE\_LC}\, I_{minE\_LC} + \theta_{minE\_HC}\, I_{minE\_HC} \tag{3.9}$$

The planning weight β_{plan} captures the influence of DV_{plan} on choice behaviour. To allow for systematic deviations from behaviour prescribed by DV_{plan}, we also included preference parameters θ. These preference parameters simply model a participant's tendency to generally choose the accept (or the reject option). The parameter θ_{basic} captures the preference in trials where participants had enough energy to accept and did not reach the maximum energy level (termed basic trials). We implemented this with a binary indicator variable I_{basic} that equals one if the current trial was basic and zero if not. To model behaviour in trials with maximum or insufficient energy, we also included three bias parameters θ_{maxE}, θ_{minE_HC} and θ_{minE_LC}. The first of these bias parameters θ_{maxE} models the special case when participants had to choose on a trial with full energy. We expect this bias parameter to be generally positive because a further reject choice would not increase the energy further. For the other two bias parameters θ_{minE_HC} and θ_{minE_LC} we expect these to be generally negative, i.e. participants will reject an offer if they have insufficient energy. Subsets of these low- and max-energy trials are again selected by an appropriate binary indicator variable.

Simple strategy model (SM). Since forward planning or other elaborate anticipatory schemes might incur considerable computational costs, participants may use a simple strategy, where action selection is only based on offer value. We define the decision variable for the SM as offer value centred across the four offer values 1 to 4:

$$DV_{simple} = o - 2.5 \tag{3.10}$$

The probability to accept is defined in the same way as for the PM

$$p(accept) = \frac{1}{1 + e^{-\eta}} \tag{3.11}$$

where

$$\begin{aligned}\eta = \beta_{simple}\, DV_{simple}\, I_{basic} + \theta_{basic}\, I_{basic} + \theta_{maxE}\, I_{maxE} + \theta_{minE\_LC}\, I_{minE\_LC} \\ + \theta_{minE\_HC}\, I_{minE\_HC}\end{aligned} \tag{3.12}$$

Here, the parameter β_{simple} captures the influence of offer value on choice behaviour.

Hybrid strategy model (HM). To cover the case that participants may choose based on both, expected long-term values and offer specific preferences, we use a hybrid strategy as a mixture of both planning and simple strategy. Such a hybrid strategy enables the decision maker to still use forward planning but mix this decision tendency with a simple strategy for each of the four offers. Note that we do not explicitly model arbitration and cannot identify which strategy dominates at any given time. However, the model enables us to test whether there is a mix of a simple and a planning strategy across trials. Like in the PM, DV_{plan} is defined as the difference between the optimal state-action values

$$DV_{plan} = Q_t(s, a = accept) - Q_t(s, a = reject) \tag{3.13}$$

The probability to accept is defined as

$$p(accept) = \frac{1}{1 + e^{-\eta}} \tag{3.14}$$

Where now

$$\begin{gathered}\eta = \beta_{plan}\, DV_{plan} + \theta_{O1}\, I_{O1} + \theta_{O2}\, I_{O2} + \theta_{O3}\, I_{O3} + \theta_{O4}\, I_{O4} + \theta_{maxE}\, I_{maxE} \\ + \theta_{minE\_LC}\, I_{minE\_LC} + \theta_{minE\_HC}\, I_{minE\_HC}\end{gathered} \tag{3.15}$$

In addition to the planning weight β_{plan} and the three bias parameters for extreme energy cases, the HM adds, as compared to the PM, four offer-specific preference parameters ($\theta_{O1}, \theta_{O2}, \theta_{O3}, \theta_{O4}$). The indicator variables ($I_{O1}, \mathrm{I}_{O2}, I_{O3}, I_{O4}$) equal one if a specific offer was presented for basic trials (i.e. energy was neither at maximum nor too low to accept). In other words, in contrast to the PM, the four offer-specific bias parameters will indicate a relative dependence on the simple strategy. For example, a negative offer-specific parameter will indicate a participants' preference to reject that specific offer.

3.3.5 Model fitting and evaluation

3.3.5.1 Model fitting.

Using a hierarchical Bayesian approach, we jointly estimated both participant- and group-level parameters. For the PM and the SM, β and θ_{basic} were allowed to vary by participant. For the HM β_{plan}, θ_{O1}, θ_{O2}, θ_{O3} and θ_{O4} were allowed to vary by participant. The parameters θ_{minE_LC}, θ_{minE_HC} and θ_{maxE} were modelled as constant over participant. The participant parameters were drawn from a normal distribution with respective group parameters μ and σ. These group parameters were themselves modelled as draws from a weakly informative hyperprior distribution: $\mu \sim Normal(0{,}2)$ and $\sigma \sim HalfNormal(0{,}2)$. A complete description of the models as

Stan code can be found online (https://doi.org/10.5281/zenodo.5112965). We fitted models using Hamiltonian Markov Chain Monte Carlo as implemented in Stan (Carpenter et al., 2017) via the PyStan interface (Stan Development Team, 2018, Version 2.19.1.1). We obtained 4,000 samples from four chains of length 2,000 (1,000 warmup) from the posterior distribution over model parameters. The potential scale reduction factor on split chains $\hat{R}$ was calculated (Gelman & Rubin, 1992), indicating convergence for all parameters ($\hat{R} \approx 1$).

3.3.5.2 Model comparison.

We compared the predictive accuracy of the PM, SM and HM using leave-one-out cross-validation approximated by Pareto-smoothed importance sampling (PSIS-LOO) (Vehtari et al., 2017) as implemented in the python package ArViz (Kumar et al., 2019, Version 0.9.0). We obtained the expected log pointwise predictive density (elpd) and its standard error on the deviance scale (-2* elpd) and refer to this quantity as leave-one-out cross-validation information criterion (LOOIC). Lower values of LOOIC indicate better model fit.

3.3.5.3 Posterior predictions.

To further assess whether the fitted models capture the observed behavioural pattern, we conducted posterior predictive checks using mixed predictive replication for hierarchical models (Gelman et al., 1996). To compute predictive replications we first sampled the group parameters (μ and σ) from the posterior and then sampled forty normally distributed participant-level parameters from these group parameters. Replicated accept-reject responses were generated for replicated participants and all trials by sampling from a Bernoulli distribution $response\_rep \sim B(p)$, where p is the response probability as defined in equations 3.8 and 3.9 (PM) or 3.12 (SM) or 3.15 (HM). The resulting array of replicated responses is of size $M = n_{samples} \times n_{participnts} \times n_{trials}$. Individual dots in Fig 3.2 A indicate choice proportions for 40 replicated participants averaged over 100 posterior samples. Bars represent averages over replicated participants and 100 samples, with error bars indicating between participant standard deviation. For the additional informal model validation in Fig 3.2 C, we computed the proportion of replicated responses that matched the participant responses. Note that the set of stimulus configurations (energy, current segment type, future segment type, trial within segment) that participants visited during the task was identical to the stimulus configurations for which replicated responses were sampled under the three models (PM, SM, HM). A match was counted (match = 1, mismatch = 0) if the model's response (accept or reject) was identical to a participant's choice in a given trial.

3.3.5.4 Parameter recovery.

To ensure that our model parameters are identifiable, we performed a parameter recovery analysis with the most complex model HM. We first generated simulated data using posterior means of the participant level parameters β_{plan}, θ_{O1}, θ_{O2}, θ_{O3} and θ_{O4} and group level parameters θ_{minE_LC}, θ_{minE_HC} and θ_{maxE}. Next, we refitted the model to the simulated data and compared the estimated parameters to the known data-generating parameters. We considered a known parameter recovered if its value was within the 95% posterior credibility interval (CI) of the re-estimated parameter. Results showed that both, participant level parameters (>99%) and all group level parameters, including those for β_{plan}, θ_{O1}, θ_{O2}, θ_{O3} and θ_{O4} could be reliably recovered from the simulated data (for details see Jupyter Notebook at https://doi.org/10.5281/zenodo.5112965).

3.3.6 Conflict and response time analysis

A key quantity for our analysis of response times (RT) and fMRI data was conflict.

$$conflict = -|Q_t(s, a = accept) - Q_t(s, a = reject)| = -|DV_{plan}| \tag{3.16}$$

This corresponds to the similarity between long-term values for accepting and rejecting (see equation 3.7). If, for a given trial and task state, the action-value difference is small, conflict is large. Conversely, if the action-value difference is large, conflict is small. We consider conflict as a signal of choice difficulty, reflecting the need for elaborate information processing such as planning. We assume that participants do not calculate the conflict directly (which would require planning by itself), but that they have quick and frugal access to a proxy for the conflict (D. G. Lee & Daunizeau, 2021).

We analysed response times using hierarchical Bayesian linear regression estimating group- and participant-level parameters simultaneously. We modelled log RT as the linear function

$$\log RT = \beta_0 + \beta_{conflict} C + \beta_{type} I + \beta_{interaction} CI \tag{3.17}$$

where C is conflict (Eq. 3.16), I is a binary indicator variable that equals one if the current offer was 2 or 3 (which we call in the following intermediate) and zero if the current offer is 1 or 4 (which we call in the following extreme). This classification into intermediate and extreme offers was based on participants' choice behaviour (Fig 3.2, A). CI models the interaction between offer type and conflict. The participant-level intercept β_0 and parameters $\beta_{conflict}$, β_{type} and $\beta_{interaction}$ were normally distributed with group parameters μ and σ. We gave these group parameters a weakly informative hyperprior: $\mu \sim Normal(0,10)$ and $\sigma \sim HalfNormal(0,10)$. Models were fit in Stan via PyStan using Hamiltonian Markov Chain Monte Carlo. We obtained 2,000 posterior samples from four chains of length 1,000 (500 warmup). The potential scale reduction factor on split chains $\hat{R}$ was calculated, indicating convergence for all parameters ($\hat{R} \approx 1$). We generated linear predictions of log RT using 2,000 posterior samples of the group hyper parameters (μ_0, $\mu_{conflict}$, μ_{type}, $\mu_{interaction}$) and exponentiated back to the original RT scale for better interpretability. The regression lines in Fig 3.3 C correspond to the median across samples and shaded areas to the 95% interval. All trials that were not timed out (RT > 5s) were included in the analysis.

3.3.7 fMRI acquisition and preprocessing

fMRI data were acquired on a 3 T MRI scanner (Siemens Magnetom Trio Tim, Siemens Medical Solutions, Erlangen, Germany) using a 32 channel head coil. On average, per participant, 942 volumes were acquired across three sessions, using a T2*-weighted echo-planar sequence (TR=2360 ms, TE=25 ms, flip angle=80°, FoV=192 mm). For each image, 48 axial slices of 2.5 mm were sampled in descending order. Field maps were acquired after each functional session (TR=532 ms, short TE=5.32 ms, long TE=7.78 ms). Structural data were acquired using a T1-weighted MPRAGE sequence (TR = 2400 ms, TE=2.19 ms, flip angle=8°, FoV=272 mm).

fMRI data were preprocessed and analysed using Statistical Parametric Mapping (SPM12) (Wellcome Trust Centre for Neuroimaging, London, UK). Functional images were unwarped using individual field maps generated by SPM's field map toolbox (Hutton et al., 2002), slice time corrected, realigned to the first image of the session, spatially normalized to the MNI template using the unified segmentation approach (Ashburner & Friston, 2005) and smoothed with an 8mm full-width at half-max (FWHM) Gaussian kernel.

3.3.8 fMRI Analysis

For each participant we specified and estimated a general linear model (GLM). Motivated by our behavioural results, we included one event regressor with response phase (see Fig 3.1, A) onsets for intermediate (2 and 3) offers and one event regressor with response phase onsets for extreme (1 and 4) offers. For each of these two event regressors, we included a parametrically modulated regressor with trial-wise conflict values. According to SPM's default orthogonalisation setting, conflict values were mean-centred per condition. Additional regressors of no interest were included: an event regressor with response phase onsets for extreme energy trials (either maximum or insufficient energy) and an event regressor with response phase onsets of accept choices (to control for the effect of action). Onsets were modelled as stick function with duration 0. Regressors were convolved with the canonical HRF. We also included the 6 movement parameter vectors as nuisance regressors. Images of all three sessions were analysed together and a constant for each session was included in the design matrix. For each participant the following first level contrasts were computed: intermediate > extreme, extreme > intermediate, conflict_intermediate > conflict_extreme and a parametric effect of conflict averaged across intermediate and extreme trials. Finally, we performed one-sample t-tests on the contrast images of all participants to assess statistical significance on the group-level. Statistical parametric maps were initially thresholded with p = 0.001 (see Tables S3.2-3.5). Voxels with a family-wise-error corrected p-value < 0.05 were considered significant.

dACC and dorsolateral prefrontal cortex (dlPFC) are commonly associated with conflict processing (E. K. Miller & Cohen, 2001). Therefore, we defined a priori regions of interest (ROI), which we used for small volume correction (SVC). ROIs were defined using the WFU-Pickatlas software (Maldjian et al., 2003) with a dilation factor of 1. The ROI for the dACC encompassed dorsal Brodmann area (BA) 32, clipped at z = 18 in MNI-space (S3.1 Fig). The ROI for the dlPFC encompassed BA46 and BA9. Generated masks are available at https://doi.org/10.5281/zenodo.5112965.

3.4 Results

3.4.1 Behavioural results

3.4.1.1 Choice behaviour.

We first identified situations in the task that could be classified as generally difficult or easy based on participants' choice frequencies. An often repeated choice pattern indicates that a specific situation can be handled by simple response mechanisms, while a mixed response pattern of accept and reject indicates that more elaborate information processing may be required. Analysis of choice frequencies revealed an obvious pattern, showing that participants accepted offer 1 in only a few trials (mean = 1%, SD = 2%) and conversely accepted offer 4 in the majority of trials (mean = 98%, SD = 3%) (Fig 3.2, A). For offers 2 (mean = 14 %, SD = 12 %) and 3 (mean = 77 %, SD = 13 %), the choice behaviour was more balanced between accepting and rejecting. To further quantify the balance between accepting and rejecting, we computed the distance between choice frequencies and the 50% chance level and compared these distances across offer values. Distances from chance level were larger for offer 1 and 4 compared to offer 2 and 3 (pairwise Wilcoxon signed-rank tests, p < 0.001). There was no significant difference between offer 1 and offer 4 (Wilcoxon signed-rank test, p = 0.094). We also found that the distance from chance level was greater for offer 2 than for offer 3 (Wilcoxon signed-rank test, p < 0.001).

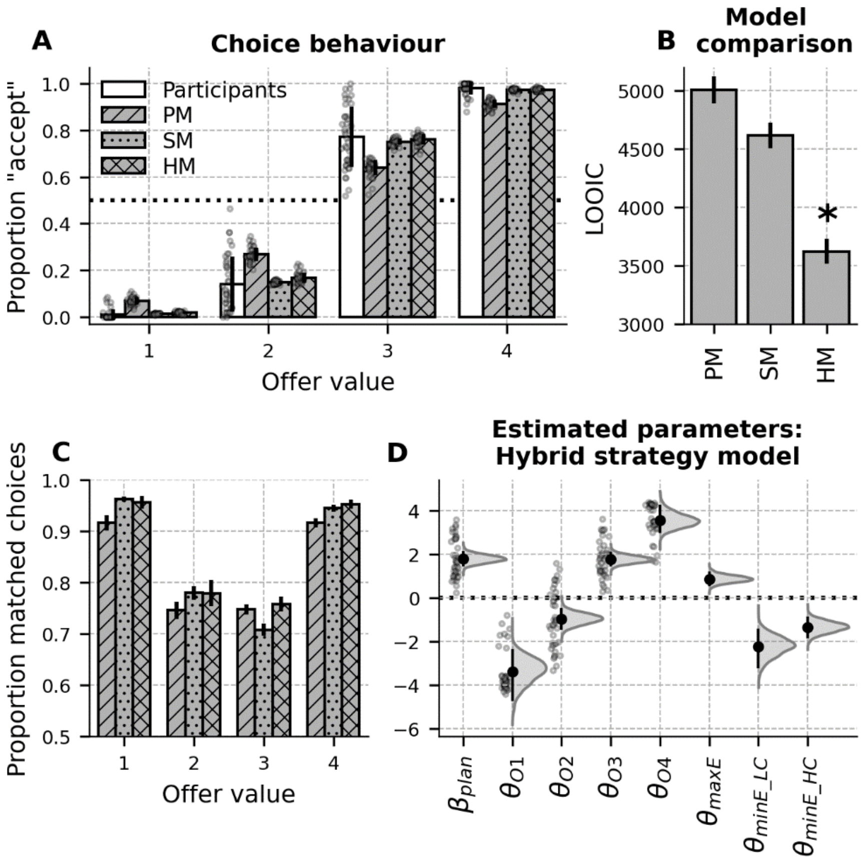


Fig 3.2. Choice behaviour and modelling results. (A) Plot of accept choices across offer values for participants and posterior predictions for the Planning strategy model (PM), simple strategy model (SM) and hybrid strategy model (HM). Bars represent the mean acceptance frequency across participants or across simulated participants. Trials in which energy was at max or too low to accept were excluded. Error bars represent standard deviation (SD). Dots represent individual participant data. The dotted line represents 50% chance level. **(B)** Model comparison of the PM, SM and HM. Error bars represent standard errors (SE) of the LOOIC. The asterisk indicates the winning model. **(C)** Proportion of simulated accept-reject choices that matched participant choices for the three models PM, SM and HM. Bars indicate averages over 200 posterior samples. The bar order and pattern is the same as in (A). Error bars represent SD. **(D)** Estimated parameters of the winning hybrid strategy model. Large black dots represent posterior means of group parameters with error bars depicting 95% credibility intervals. Grey curves represent kernel density estimates for the posterior distributions of group parameters. Semi-transparent small dots represent posterior means of participant-level parameter estimates.

From this pattern, we hypothesised that participants might have treated the choice given an extreme offer 1 or 4 as generally easy and the choice given an intermediate offer 2 or 3 as generally hard. We hypothesized that this categorisation into what we call control contexts predetermines the actual planning investment in a given trial. In the following we will test this hypothesis and provide further insights using model-based analysis of choice behaviour, analysis of response times and analysis of fMRI BOLD-signals.

In addition to the offer value, we found, using logistic regression, that participants' choice behaviour was also influenced by other task features (S3.1 Table). Participants chose more often the accept option if they had more energy units and likewise if the current energy cost of accepting was low (current segment = LC). Participants chose more often the accept option if the upcoming energy cost was high (future segment = HC), showing that participants considered information about the future segment when making a decision.

3.4.1.2 A combination of forward planning and simple offer specific preferences fitted behaviour best.

We assumed that our task design motivated participants to use simple heuristics and forward planning in a situation-appropriate way. To test for the simultaneous presence of planning and simple heuristics we carried out a model-based analysis of choice behaviour. Three different strategies of how participants select their responses were considered. First, participants may have fully relied on forward planning across the current and the next segment to select actions that maximize the expected value (PM). Second, participants may have used a simple strategy just dependent on the offer value (SM). To illustrate the difference between these two strategies, let us consider a (for convenience deterministic) agent in the first trial of a segment, with three energy units, where the current and the future energy cost is 2 (segment pair HC/HC) and offer 3 is presented. A planning agent would reject the offer and replenish its energy reserves in order to be able to accept potential better offers in the future. In contrast, an agent following a simple strategy, e.g. who always accepts offers 3 and 4 and rejects offer 1 and 2, would accept the offer 3. We also considered a third alternative that participants use a mixture between planning and a simple strategy (HM) to achieve a good trade-off between the benefits and costs of the respective strategies depending on the current task situation.

We compared how well the three cognitive computational models fitted participants' behaviour using leave-one-out cross validation. We found, as shown in Fig 3.2 B, that the HM explained participant behaviour substantially better (LOOIC = 3622.5, SE = 108) than the PM (LOOIC = 5006.0, SE = 117.5) and the SM (LOOIC = 4613.2, SE = 109.4). This demonstrates that participants use both planning and simple heuristics throughout the task. To confirm this result, we also compared models on the participant level and found that the HM explained behaviour best for 33 (of N = 40) participants (S3.2 Fig). For 3 participants the SM and for 4 participants the PM explained behaviour best.

We also simulated posterior predictions for the three models and plotted acceptance frequencies across offer values (Fig 3.2, A). Both, the HM and the SM, closely captured the behavioural pattern of participants, but acceptance frequencies of the PM were lower for offer 1 and 2 and higher for offer 2 and 3. These simulations are consistent with participants mixing forward planning with a simple reject-preference for offers 1 and 2 and an accept-preference for offers 3 and 4. As a further informal illustration of why the HM was superior to the SM, we computed the proportion of matches between participant choices and the simulated choices from the fitted models (Fig 3.2, C). While SM shows high matching rates for offers 1, 2 and 4, the matching rate for offer 3 is decisively lower compared to the HM. Conversely, the PM has considerable lower matching rates than the HM for offers 1,2,4 but achieves a relatively high matching rate for offer 3. These results show that the SM particularly fails to account for participants choices for offer 3, presumably because participants engage in an increased amount of planning for offer 3 (see Fig 3.2A, where the mean accept rate for offer 3 is closest to the 50% line among all four offers, i.e. offer 3 does not support a simple action selection strategy).

Parameter estimates of the HM demonstrate both evidence for forward planning and usage of a simple strategy as quantified by four offer-specific preferences (Fig 3.2, D). We found a positive weight on the planned value difference (mean group parameter β_{plan} = 1.79, 95% CI = [1.45, 2.15]). This indicates that participants, when making a decision, accounted for its future consequences. We also found preference parameters different from zero for all four offers. For offer one (mean group parameter θ_{O1} = -3.39, 95% CI = [-4.76, -2.37]) and offer two (mean group parameter θ_{O2} = -0.96, 95% CI = [-1.47, -0.47]), participants showed a preference for rejecting (indicated by negative parameter values). For offer three (mean group parameter θ_{O3} = 1.77, 95% CI = [1.43, 2.11]) and offer four (mean group parameter θ_{O4} = 3.56, 95% CI = [2.97, 4.28]), participants showed a preference for accepting (indicated by positive parameter values). We explicitly modelled special cases, where participants had either maximum energy (θ_{maxE}) or not enough energy to accept (θ_{minE_LC} and θ_{minE_HC}), see also Methods. As expected, participants showed a bias to accept in maximum energy trials and a bias to reject in low energy trials (see Fig 3.2, D).

We further found evidence that participants who account for the long-term consequences of their actions by e.g. planning, earn more points. Post-hoc correlation analysis revealed that participants with a larger fitted planning parameter β_{plan} of the winning hybrid model accumulated more points throughout the experiment (r = 0.521, p = 0.001, Fig 3.3, A) and had slower average response times (r = 0.374, p = 0.017, Fig 3.3, B).

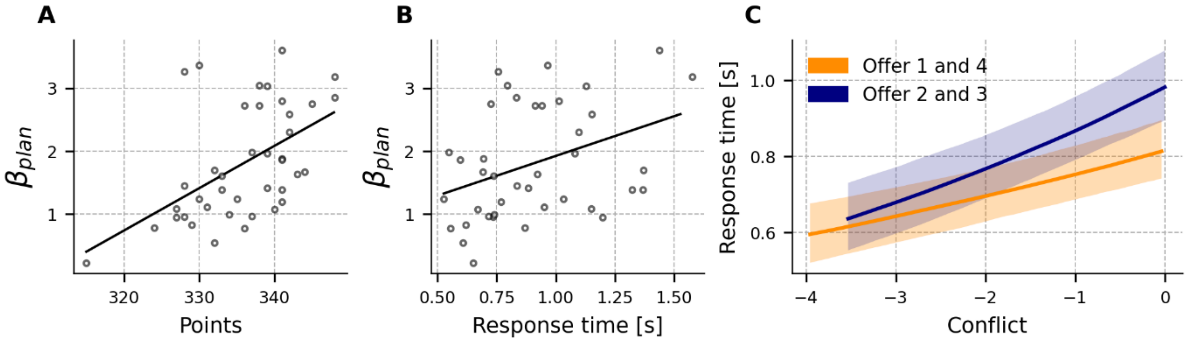


Fig 3.3. Correlation of fitted planning weight with behavioural markers and response time analysis. (A) Participant level posterior mean planning weight of the winning hybrid model versus accumulated points **(B)** Participant level posterior mean planning weight of the winning hybrid model versus mean response times. **(C)** Linear regression of log response times against choice conflict. Larger choice conflict was associated with an increase in response time, where the increase was significantly more pronounced during intermediate offers 2 and 3 compared to the extreme offers 1 and 4. Regression lines were computed from mean group hyperparameters. Shaded areas depict 95% credibility areas computed from 2,000 posterior samples of the group hyperparameters.

3.4.1.3 Greater conflict-driven increase of response times in intermediate than extreme trials.

Previous research suggested that the brain regulates the use of cognitive control based on the estimated value of control (Shenhav et al., 2013). Analogously, we assume that the brain uses similar value estimates when deciding about the degree of forward planning during sequential decision making. Most importantly we hypothesized that, due to cost incurred by the computation of control values themselves, the brain uses a context-specific prior assumption about the general need for planning to minimize the "metacosts" of control decisions. To further test this hypothesis we analysed the relationship between response times as an indicator for the degree of planning and planning value (operationalised by a specific conflict measure, see methods). We expected that not only would larger conflicts generally lead to increased response times, but critically that this increase will be more pronounced for the intermediate offers 2 and 3, possibly reflecting context-specific planning activity driven by a context-specific evaluation of conflict.

Bayesian linear regression indeed showed that conflict was significantly more predictive for log RT for the intermediate offers 2 and 3 compared to the extreme offers 1 and 4 (group parameter $\beta_{interaction}$ = 0.04, 95% CI = [0.02 0.07]). We also found a significant positive main effect of conflict (group parameter $\beta_{conflict}$ = 0.08, 95% CI = [0.05 0.11]) and offer type (group parameter $\beta_{is\_intermediate}$ = 0.18, 95% CI = [0.14 0.23]) on log RT. Fig 3.3 C shows fitted regression lines on the untransformed RT scale. This shows that, the increase in computation time associated with conflict level is greater for intermediate offers than extreme offers.

3.4.2 FMRI results

3.4.2.1 A frontal network is more activated in intermediate than extreme trials.

A set of parietal and frontal regions, sometimes called the multiple demand network, are often activated during cognitively challenging tasks (Duncan, 2010; E. K. Miller & Cohen, 2001). To relate to these well-established findings, we first tested to confirm where brain activity was higher during intermediate compared to extreme offers (Intermediate > Extreme) expecting to see greater activity in frontal areas related to planning and cognitive control. We indeed found significantly greater activity in dACC and right dlPFC (Fig 3.4, A; Table 3.1 and S3.2 Table).

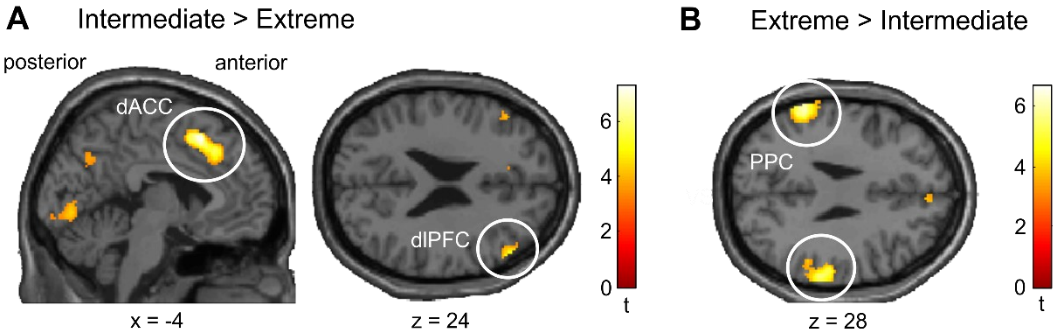


Fig 3.4. Different activation of dACC for intermediate versus extreme offers. (A) dACC and right dlPFC were significantly more activated during intermediate compared to extreme offers. **(B)** Bilateral PPC was significantly more activated during extreme compared to intermediate offers. Activations displayed at p < 0.001 uncorrected. See Table 3.1 for peak MNI-coordinates and statistics, significant at p<0.05 FWE corrected.

Table 3.1. Summary of fMRI Results

	t	FWE p value		MNI coordinates		
Region		whole brain	SVC	x	y	z
Intermediate > Extreme						
dACC	7.26	< 0.001	< 0.001	-4	16	52
dlPFC R	5.12	0.180	0.018	52	34	24
Occipital L	6.98	< 0.001		-18	-92	-10
Occipital R	6.38	0.007		20	-84	-8
Extreme > Intermediate						
PPC (BA 40) L	6.68	0.003		-60	-44	38
PPC (BA 40) R	5.49	0.072		64	-32	28
Conflict						
dACC	4.82	0.290	0.007	10	24	44
anterior Insula L	5.62	0.041		-30	24	-2
anterior Insula R	6.50	0.004		38	20	0
Conflict: Intermediate > Extreme						
dACC	4.32	0.744	0.030	8	2	40
posterior MTG R	5.69	0.040		56	-64	6

Regions significant (p < 0.05 FWE corrected) at the whole brain level or small volume correction (SVC) based on a priori regions of interest. Only clusters with more than 10 voxels were reported. BA, Brodmann area; L, left; R, right; dACC, dorsal anterior cingulate cortex; dlPFC, dorsolateral prefrontal cortex; Occipital, occipital lobe; VS, ventral Striatum; PPC, posterior parietal cortex; MTG, middle temporal gyrus.

We also tested where brain activity was greater during extreme versus intermediate trials (Extreme > Immediate) and found increased activity in bilateral posterior parietal cortex, where the cluster in the left hemisphere was significant at the whole brain corrected level (PPC; Fig 3.4, B; Table 3.1 and S3.3 Table). Besides its role in sensory attenuation, posterior parietal cortex is also involved in sensorimotor transformations during decision making (Andersen & Cui, 2009). This suggests that participant decisions for extreme offers might be related to low-level sensorimotor processes, coupling simple stimulus cues to actions. A network including left ventral Striatum (VS), posterior cingulate cortex (PCC) and bilateral Amygdala emerged at a lower threshold (see S3.3 Table). These regions have been shown to encode value information during reward-based choice (Bartra et al., 2013). Higher activation in these areas during extreme trials might indicate an increased salience of offer value information instigating a simple response strategy based on offer-specific preferences. However, this idea requires further research.

3.4.2.2 Context-dependent conflict processing in dACC.

As a confirmatory analysis of previous findings implicating the dACC in the monitoring of various signals to evaluate the need for additional control (e.g. conflict, Shenhav et al., 2013), we also tested for the effect of conflict averaged across conditions. In accord with this previous research, we found a significant positive correlation with BOLD-activity in the dACC (Fig 3.5, A; Table 3.1 and S3.4 Table). We also found a significant positive effect of conflict in bilateral anterior Insula (Fig 3.5, A; Table 3.1 and S3.4 Table).

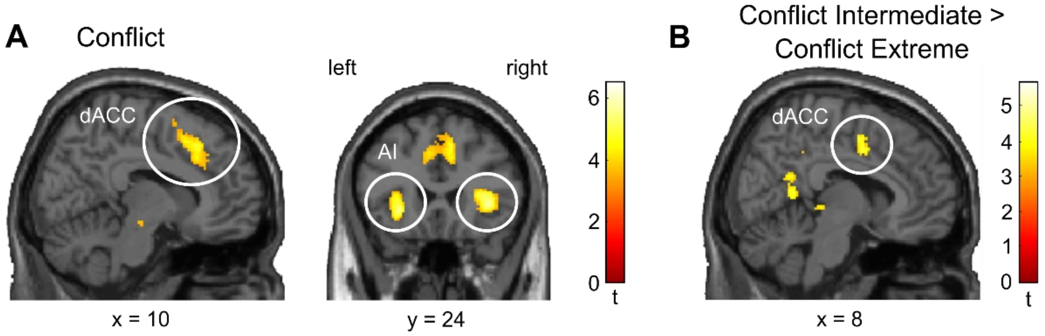


Fig 3.5. Effects of conflict. (A) BOLD-activity in dACC and bilateral anterior Insula (AI) correlated with conflict. **(B)** The correlation of BOLD-activity in dACC with conflict was significantly increased during intermediate versus extreme offers. Activations displayed at p < 0.001 uncorrected. See Table 3.1 for peak MNI-coordinates and statistics, significant at p<0.05 FWE corrected.

Next, we tested our main hypothesis that the extreme and intermediate conditions are treated as different control contexts by the brain. Note that one obvious reason for the effects in the categorical contrast Intermediate > Extreme could be that average conflict was higher in intermediate trials than in extreme trials (S3.3 Fig). Another possibility would be that the brain areas involved in processing conflict are modulated by the context. In other words, is conflict processed differently in the brain when the subject is in an intermediate compared to an extreme trial? Behavioural results in Fig 3.3 C already indicate such a context-dependent mechanism, showing that reaction times increased more with conflict during intermediate than extreme trials.

To test this context dependency using brain activity we included conflict as parametric modulator in our GLM, separately for intermediate and extreme offers. We then computed a contrast between the parametric modulator of conflict for intermediate offers minus the parametric modulator of conflict for extreme offers (Conflict Intermediate > Conflict Extreme). Our expectation was that the dACC would track conflicts (as a proxy for the value of planning), but to a lesser extent in a context with a low prior need for planning (i.e. in the extreme context), due to the metacosts associated with obtaining conflict values. We indeed found that BOLD-activity in dACC and right posterior middle temporal gyrus (pMTG) was more strongly correlated with conflict during intermediate offers compared to extreme offers (Fig 3.5, B; Table 3.1 and S3.5 Table). An effect in dlPFC emerged at a lower threshold (see S3.5 Table). This finding aligns well with the results of the reaction time analysis and is consistent with the idea that the situation-appropriate investment into planning is driven by a context-dependent evaluation of conflict involving the dACC.

3.5 Discussion

We used a novel sequential task with a complex task space to investigate how people decide when to plan ahead. We found evidence that participants use readily available features of the task space, such as offer values, to construct contexts that condition the balancing between forward planning and a simpler response strategy. We further provided evidence that the context dependency of planning might be mediated by context-dependent conflict processing involving dACC. Our study provides initial evidence that the human ability to efficiently allocate cognitive control in complex tasks is supported by category-based cognition that harnesses regularities in control demand to generate control contexts.

Normatively, a decision about the engagement into elaborate planning should find the optimal trade-off between the benefits and costs of such planning in a given situation (Shenhav et al., 2013). It is however unlikely that people calculate the expected value of planning explicitly because this would require planning itself. Previous studies have therefore argued that people must have quick and automatic access to a proxy of the value of planning (D. G. Lee & Daunizeau, 2021; Lieder et al., 2018). It is a largely unresolved question how humans learn such proxies to decide about their engagement into effortful planning. A hallmark of learning is that people generalize individual sensorimotor experiences into broader categories allowing them to react adequately to novel instances of a learned category (Konidaris, 2019; E. E. Smith & Medin, 1981; Tenenbaum et al., 2011; Yee, 2019). Recent research suggests that the brain leverages similar generalisation mechanisms when deciding about the allocation of cognitive control or when computing meta-control (Bhandari et al., 2017; M. Botvinick et al., 2020; Lieder et al., 2018; Marković et al., 2020; Schwöbel et al., 2021). Here, we tested this principle using a complex planning task, where the demand for planning changed depending on the current situation as defined by a configuration of task features including offer value, energy, current energy cost, future energy cost and trial within a segment.

Our results suggest that participants used a generalized task representation, mapping clusters of states (contexts) to an approximated value of planning (e.g. conflict) to decide efficiently about the investment in planning. In particular, our results suggest that participants constructed two contexts, one for the extreme offers 1 and 4, associated with a low prior tendency for planning, and one for the intermediate offers 2 and 3 associated with a high prior tendency for planning. Three pieces of converging evidence supported this conclusion: First, responses in an intermediate context were more mixed between accepting and rejecting compared to an extreme context. Second, our model-based analysis suggested that the mixed response profile for the intermediate offers can be explained by participants planning ahead multiple steps into the future (especially offer 3 ; see Fig 3.2; C). Third, the findings that the correlation of conflict with response times and dACC activity depended on the current context further corroborated that participants had learned an offer-specific context structure.

To investigate the usage of forward planning or simple heuristics and its determining factors, we designed a task where both types of decision making can occur. Our computational analysis indeed revealed that a model including both, planning and simple offer-specific preferences, fitted behaviour substantially better, than a model that relied on planning or a simple strategy alone. Our model implemented planning as a search through a decision tree, calculating expected long-term state-action values. However, due to the high computational cost of such an exhaustive search, it is unlikely that participants planned in exactly this way. We therefore assumed that participants used some sort of approximate planning, reducing planning complexity while still accommodating for an action’s future consequences. Pruning of decision trees (Huys et al., 2012), adjusting the planning depth (Keramati et al., 2016) or the division of a decision problem into smaller subproblems (Huys et al., 2015) are just some ways of how people can adjust planning complexity. While we cannot say anything about the exact implementation of planning, we found that the fitted individual planning weights of the winning hybrid model correlated with RT, suggesting that participants engaged in some kind of forward planning. Another limitation of the hybrid model is that it does not explicitly model arbitration between different decision modes, but only recognises the presence of a mixture. While research on model-free and model-based reinforcement learning systems provided important insights into the neurocomputational mechanism of arbitration using a simple (two-step) planning task (Kool et al., 2017; S. W. Lee et al., 2014), less is known about how the brain adapts its decision mode during more complex tasks. Our study provides evidence for a link between the generalisation of task representations and high-level control decisions and may therefore inform future attempts to model arbitration mechanisms in more complex realistic environments.

Our finding that BOLD activity in the dACC correlated with conflict is consistent with the view that the dACC monitors the need for effortful controlled processing (M. M. Botvinick et al., 2001; Shenhav et al., 2013; Shenhav et al., 2017). In our task, when it was not clear at the beginning of a trial whether accepting or rejecting is the best option (i.e. if the conflict was high), participants needed to generate additional information by planning ahead which could then help to adjudicate between the two options. While dACC might have played a central role in detecting the need for additional planning, a distributed network including dlPFC and other structures might have been involved in the additional information sampling. The finding that dlPFC was more active in the demanding intermediate context is consistent with the view that dlPFC is central to planning (Fuster, 2015). In order for planning processes to have an impact on the decision, it is often necessary to inhibit a prepotent response first. We found evidence that such prepotent responses played a role in in our task as well, as our model-based analysis revealed non-negative choice preference parameters for all offers. Previous research suggests that such prepotent responses could have been inhibited by a hyper-direct pathway from the ACC to the basal ganglia, effectively increasing time until movement generation and thus allowing planning structures to influence decision making (Cavanagh et al., 2011; Gluth et al., 2012; Wiecki & Frank, 2013; Wiecki et al., 2013).

The anterior insula (AI) is often coactivated with the ACC in classical cognitive control tasks (Duncan & Owen, 2000) and sequential tasks alike (e.g. Schwartenbeck et al., 2015). Previous research suggests that, besides of its role in interoception and general awareness (Craig & Craig, 2009), the AI appears to be specifically involved in the representation and learning of uncertainty (Bossaerts, 2010; Loued-Khenissi et al., 2020; Singer et al., 2009). Translated to our task, the AI could have relayed information about uncertainty (or conflict) to the dACC, which then initiates adaptive behavioural change in the form of planning.

We found that the correlation of activity in the dACC with conflict (which we take to be a proxy for the value to plan ahead) depended on context. One possible explanation for this pattern could be that the dACC has access to a hierarchical representation of learned conflicts, whereby conflicts encoded at a finer level of task space are subsumed under conflicts encoded at the level of context. In other words, states of similar difficulty could be grouped into a more general category that e.g. simply indicates whether the decision is easy or difficult. In contexts with a high prior expectation of conflict, i.e. in an intermediate context, the dACC could access conflict at a more fine-grained level to enable the appropriate level of planning. Conversely, in a context with low prior expectation of conflict, i.e. in an extreme context, the dACC would not access information beyond that at the coarse context level, as the overall need for planning was low anyway. Speculating on the algorithmic implementation of such a process, the context-dependent prior assumption about conflict could set the threshold for the meta-decision problem of inferring the need of planning. In an intermediate context, a high meta-threshold would grant enough time for a state-level readout of conflict, whereas in an extreme context the need for planning would have been determined before state-level conflicts were accessed. We also found evidence that right posterior middle temporal gyrus (pMTG) is more correlated with conflict in an intermediate than in an extreme context. Previous research implicated the pMTG in category-based cognition (Martin, 2007). It is therefore an intriguing possibility that the pMTG is also capable of forming abstract categories of choice difficulty that support the context-dependent evaluation of planning demands. Although we can only speculate about the role of pMTG, the question how brain mechanism for structured knowledge acquisition and cognitive control interact is an important direction for future research. Overall, our findings are generally consistent with the view that people exploit the structure of a task for efficient storage and access of the value of control (Lieder et al., 2018).

3.6 Supporting information

S3.1 Table. Logistic regression of choice (accept = 1, reject = 0) against task features.

	Estimate	SE	z-value	P(>|z|)	
Intercept	-12.53	0.33	-37.58	< 10^{-20}	***
Offer value	3.65	0.09	40.80	< 10^{-20}	***
Energy	0.63	0.04	15.30	< 10^{-20}	***
Current segment (LC=1 , HC=0)	2.38	0.11	22.55	< 10^{-20}	***
Future segment (LC=1, HC=0)	-0.24	0.09	-2.75	0.00593	**
Trial	-0.02	0.04	-0.57	0.56753	

Trials were pooled across participants (N = 40). Trials in which energy was at maximum or too low to accept any offer were excluded.

S3.2 Table. fMRI results for the contrast: intermediate > extreme

	cluster size	t	FWE p value		MNI coordinates		
Region	(voxels)		whole brain	SVC	x	y	z
Dorsal anterior cingulate	929	7.26	<0.001	<0.001	-4	16	52
Occipital L	1252	6.98	<0.001		-18	-92	-10
Occipital R	1475	6.38	0.007		20	-84	-8
Dorsolateral prefrontal R	106	5.12	0.180	0.018	52	34	24
Superior parietal L	76	4.68	0.456		-28	-64	52
Putamen R	95	4.64	0.486		20	16	-4
Angular R	112	4.54	0.580		34	-70	44
Occipital L	92	4.51	0.602		-24	-84	14
Anterior Insula L	45	4.09	0.918		-26	18	0
Dorsolateral prefrontal L	30	4.06	0.935	0.223	-48	32	24
Precentral L	21	3.98	0.961		-36	-12	50
Inferior parietal R	50	3.97	0.964		36	-48	38
Superior frontal L	11	3.71	0.997		-24	40	-14
Precuneus R	128	3.69	0.998		6	-62	34

The table contains all clusters with more than 10 voxels that survived uncorrected statistical thresholding with p < 0.001.

S3.3 Table. fMRI results for the contrast: extreme > intermediate

	cluster size	t	FWE p value		MNI coordinates		
Region	(voxels)		whole brain	SVC	x	y	z
Posterior parietal L	483	6.68	0.003		-60	-44	38
Posterior parietal R	439	5.49	0.072		64	-32	28
Posterior cingulate	106	5.15	0.169		6	-6	36
Amygdala L	186	4.92	0.280		-24	-2	-18
Middle temporal L	391	4.90	0.296		-48	-64	6
Middle temporal R	237	4.86	0.323		48	-12	-12
Amygdala R	94	4.56	0.560		28	-2	-18
Ventral striatum	23	4.40	0.701		-4	22	-10
Middle temporal R	109	4.20	0.858		66	-48	10
Anterior cingulate	16	4.09	0.920		-4	32	4
Anterior Insula R	20	3.87	0.984		30	20	-16
Superior frontal R	17	3.71	0.997		10	48	26

The table contains all clusters with more than 10 voxels that survived uncorrected statistical thresholding with p < 0.001.

S3.4 Table. fMRI results for parametric effect of conflict

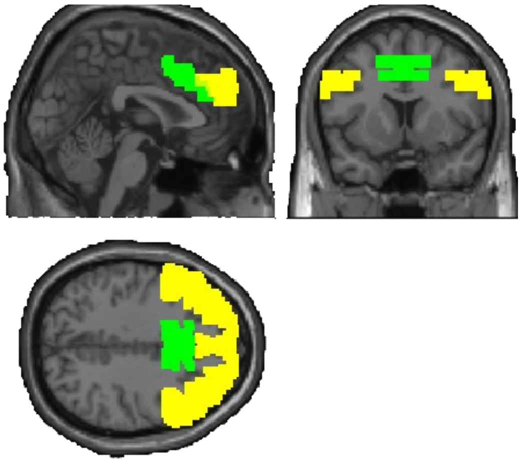
	cluster size	t	FWE p value		MNI coordinates		
Region	(voxels)		whole brain	SVC	x	y	z
Anterior Insula R	523	6.50	0.004		38	20	0
Anterior Insula L	600	5.62	0.041		-30	24	-2
Dorsal anterior cingulate	644	4.82	0.290	0.007	10	24	44
Thalamus R	40	4.29	0.718		8	-18	-12
Pallidum L	29	4.01	0.909		-12	0	-4
Supplementary Motor Area R	24	3.78	0.983		12	8	62

The table contains all clusters with more than 10 voxels that survived uncorrected statistical thresholding with p < 0.001.

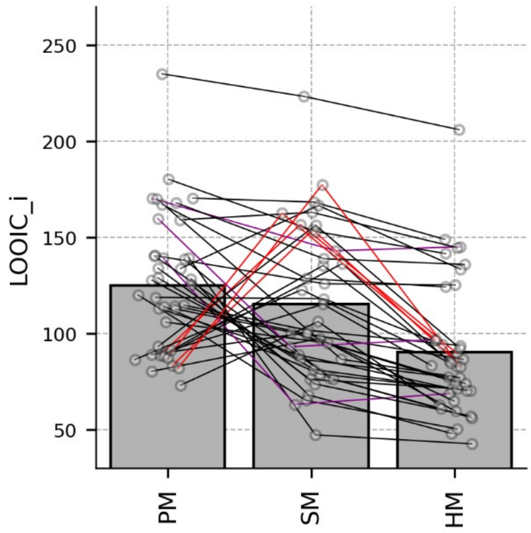
S3.5 Table. fMRI results for the contrast: conflict intermediate > conflict extreme

	cluster size	t	FWE p value		MNI coordinates		
Region	(voxels)		whole brain	SVC	x	y	z
Middle temporal R	489	5.69	0.040		56	-64	6
Anterior Insula R	389	5.05	0.196		36	-8	18
posterior cingulate L	29	4.79	0.345		-16	-32	40
Occipital L	441	4.75	0.378		-48	-74	24
Precuneus R	117	4.54	0.551		10	-50	10
Dorsolateral prefrontal R	208	4.49	0.592	0.083	54	28	6
Angular R	50	4.41	0.667		38	-62	26
Precuneus R	50	4.36	0.712		12	-44	40
Dorsal anterior cingulate	105	4.32	0.744	0.03	8	2	40
Cuneus L	91	4.28	0.774		-10	-68	26
Superior temporal R	32	4.16	0.859		46	-42	22
Thalamus L	32	4.08	0.905		-6	-30	-2
Occipital L	35	4.03	0.929		-24	-86	38
Occipital L	26	3.93	0.966		-16	-48	0
Superior temporal R	19	3.90	0.971		44	-14	-10
Precentral R	117	3.83	0.984		40	-18	58
Superior temporal L	44	3.72	0.995		-54	-20	12
Thalamus R	20	3.72	0.995		8	-30	-2
Precuneus R	60	3.71	0.996		16	-60	26
Supplementary motor area R	21	3.69	0.997		4	-8	58
Middle temporal L	12	3.69	0.997		-62	-52	0

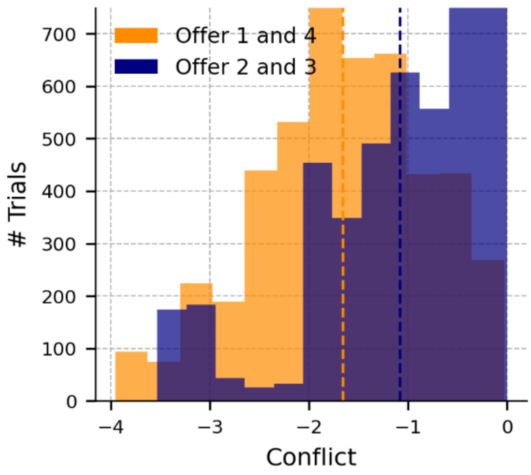
The table contains all clusters with more than 10 voxels that survived uncorrected statistical thresholding with p < 0.001.



S3.1 Fig. Definition of anatomical ROIs for small volume correction. dACC (green), dlPFC (yellow).



S3.2 Fig. Comparing models on the participant level. To calculate the predictive accuracy for each participant (LOOIC_i), pointwise log predictive densities (see methods) have to be summed up for all trials within each participant (white dots). Note that this differs from Fig 2B, in which the individual pointwise log predictive densities are summed up over all trials and participants. The hybrid strategy model (HM) was best for 33 (of N = 40) participants (black lines). For 3 participants the simple strategy model (SM, violet lines) and for 4 participants the planning model (PM, red lines) explained behaviour best. Bars indicate average LOOIC_i.



S3.3 Fig. Encountered conflict levels plotted for intermediate (2, 3) and extreme (1, 4) offers. Histograms contain data of all forty participants. Dashed lines indicate the mean.

S3.1 Text. Task instructions. Participants were guided step by step through the task on the computer screen. Written instructions were accompanied by an arrow pointing towards the stimulus element currently addressed. The written instructions were:

- Dear participant, the experiment is about collecting as many points as possible. Depending on your score at the end of the experiment, you will be paid a monetary bonus.
- The current score is indicated by the upper yellow bar.
- You can get points by accepting offers.
- The current offer is presented in the middle of the screen. The magnitude of offers varies between 1 and 4, represented by the number of golden trophies. The offers are drawn at random, having the same occurrence probability of 25%.
- However, accepting an offer is associated with energy costs. Your current energy level is represented by the lower blue bar. If you accept an offer and do not have enough energy, no points will be credited to you and the next trial will begin.
- You can replenish your energy account by selecting the "reject" option. This will increase your energy level by 1 and the next trial will begin.
- The energy level can have a maximum value of 6.
- The experiment is divided into segments, each consisting of 4 trials. Two numbers are displayed on the screen to indicate how far you are in the current segment.
- There are 2 different segment types, in which the energy costs for accepting an offer differ. In segments with 1 flash, 1 energy unit is subtracted when you accept an offer. In segments with 2 flashes, 2 energy units are subtracted when you accept an offer. The left blue-orange box at the bottom right of the screen informs you about the type of the current segment.
- In addition to the type of the current segment, information about the energy costs in the next segment is available. This can be seen in the right blue-orange box at the bottom right of the screen.
- Breaks: During the main experiment in the scanner, you have the possibility to pause twice. The pause screen is automatically displayed. You decide when you are ready to continue the experiment. Note: After a pause your score will be reset to 0. This has no effect on your final bonus. Your score is counted continuously.
- Deadline: You have a maximum of 5 seconds for each decision. If you exceed this time limit, the next trial will begin without points being awarded.
- Training: Before the main experiment in the scanner starts, you will be given a few training trials on the PC to familiarize yourself with the experiment. There is no deadline in the training and the points gained here have no effect on the bonus paid out. Please try to get as many points as possible anyway. The training phase will end automatically.

S3.1 Acknowledgments

The authors thank Florian Bolenz and Philipp T. Neukam for valuable comments and suggestions. This work was funded by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft), SFB 940 - Project number 178833530, TRR 265 - Project number 402170461 and was partially supported by Germany’s Excellence Strategy – EXC 2050/1 – Project number 390696704 – Cluster of Excellence “Centre for Tactile Internet with Human-in-the-Loop” (CeTI) of Technische Universität Dresden.

4 General discussion

4.1 Summary of key results

Planning ahead allows us to mentally simulate the effects of future actions in order to then choose the most favourable course of action. However, planning takes time and is thus associated with an opportunity cost. Normatively, one should only invest into planning if the expected benefits of planning exceeds its costs. The exact computation of those costs and benefits is infeasible, because this would incur its own computational costs (see e.g. section 1.5). Therefore, the brain can only approximate an optimal cost-benefit trade-off when regulating the use of planning. However, the cognitive and neural mechanism by which this is achieved are still largely unknown.

This dissertation tested the hypothesis that humans construct and use so-called control contexts to efficiently adjust the degree of planning to the demands of the current situation. The control context hypothesis postulates that when learning the value of planning, our brains cluster together states with similar demands for planning. This generalisation reduces the complexity of the representational space while achieving sufficiently accurate predictive coding of the value to plan ahead for a large number of future situations.

In the first study, we used a 15-step sequential decision making task to test how forward planning is dynamically mixed with simple heuristics when progressing towards a goal. When the goal was temporally distant the number of actions needed to achieve the goal was large and planning was particularly costly. However, as the goal was getting closer, planning complexity decreased and the effort to invest in it might have become worthwhile. In this scenario, we assumed that the participants identify “distance from goal” as the relevant control context for the regulation of planning. We specifically predicted that participants rely on simple heuristics when far from the goal but progressively increase their planning effort towards the end of the goal-reaching episode. To test our hypothesis, we formulated the task as a Markov Decision Process and operationalized forward planning as finding the optimal policy via dynamic programming. To assess the participants’ reliance on either forward planning or heuristics we used optimal-agent comparisons and a model-based analysis using a parametrized version of the planning model. The results confirmed our hypothesis showing that participants made suboptimal choices frequently when still far from the goal (around 40% in the first 11 trials) but rapidly approximated optimal behaviour within the last four trials. The model-based analysis corroborated that the initial suboptimality was indeed caused by participants relying on a simple heuristic (captured by the strategy preference parameter θ) and that this heuristic was progressively outweighed by forward planning (captured by the precision parameter β) when the goal was approached.

In the second study we investigated how inferred control contexts facilitate the situation-appropriate investment into forward planning via a contextually modulated processing of trial-by-trial conflicts in the dACC. To address this question we used a complex sequential decision making task, in which participants had to sustainably invest a limited but replenishable energy resource, that was needed to accept offers, in order to accumulate a maximum number of points in the long run. Importantly, the benefits of planning varied across the different situations that could be encountered in the task. Proactively adjusting the control system based on prior expectations about the demand for planning can be beneficial (see section 1.3.3), but representing such an expectation for each individual situation (448 in our task) would severely tax the memory system. We therefore assumed that participants use a more coarse-grained representation (i.e. control contexts) to precondition their final planning investment in a particular situation. We specifically assumed that inferring to be in a context, where it is generally beneficial to plan ahead, will upregulate the processing of trial-by-trial conflicts in the dACC. On the other hand, inferring to be in a context, where the demand for planning is generally low, will have an inhibitory effect on the processing of trial-by-trial conflicts in the dACC. That means in a trial in which the two options have similar values (i.e. in which conflict is high) less effort will be invested in refining those value estimates by planning ahead if participants infer to be in a context with low control demand relative to a context with high control demand.

Analysis of choice behaviour, reaction times and brain activity, provided evidence consistent with our predictions. We first tested whether participants adapt their planning to the demands of the situation or whether they consistently use either simple heuristics or a pure planning strategy, regardless of the situation. In accordance with an adaptive planning strategy, a model-based analysis of choice behaviour showed that a hybrid model that includes both forward planning and simple offer-specific preferences explained participants' behaviour best. We next tested our control context hypothesis on the behavioural level by analysing the relation of reaction times with conflict and how this relationship differs between control contexts. Here, it is assumed that reaction times reflect the time invested in planning and that conflicts (i.e. the difference between the action values derived from the model) represent preference uncertainty and thus a signal for planning demand. As predicted, we found that reaction times increased with the level of conflict, but that this increase was stronger in a context with a high demand for planning. This suggests that the amount of cognitive resources and time invested for the prospective evaluation of choice options does not only depend on the current preference-uncertainty but also on the demand for planning in the more global context. We finally tested, whether the context-dependent relationship between conflicts and forward planning is driven by context-dependent conflict processing in the dACC. Our analysis of fMRI data indeed revealed that activity in the dACC was positively correlated with trial-by-trial conflicts, but that this correlation was stronger in a context with a high demand for planning. Taken together, the results suggest that the dACC integrates representations of planning demand on different levels of abstraction to regulate prospective information sampling, i.e. forward planning, in an efficient and situation-appropriate way.

4.2 Abstraction and cognitive control

The results of this dissertation, and in particular of study 2, suggest that the abstraction of representational space plays an important role in the allocation of cognitive control. Although the role of state abstraction has received relatively little attention in research on control allocation (however for related research see Bhandari et al., 2017), it has been extensively explored in other domains such as language, perception and action control (Behrens et al., 2018; Friston, 2010; Friston & Kiebel, 2009; Lake et al., 2017; Tenenbaum et al., 2011). Integrating findings from these domains will be necessary to improve our understanding of the intertwined relationship between cognitive control and structured representations and show how control-related properties might shape structural learning.

A basic form of abstraction is already evident in visual object recognition. In every encounter with an object, the primary visual stimulus is never exactly the same, due to differences in viewing angle, illumination, occlusion, viewing distance and other factors. Fortunately however, humans are able to robustly infer the identity of objects despite of these variances. This ability relies on the hierarchically organized layers of the ventral visual stream where each cortical area abstracts away the details below its input area (DiCarlo et al., 2012). Figuratively speaking, the ability to recognize objects might be based on abstract object representations that allow to generalise novel sensory instances of that object. High-level object representations not only facilitate invariant object perception, but more importantly, they form the basis for learning appropriate behaviour. For example if one encounters a tiger that looks slightly different from the tigers seen before, on could infer, based on generalized previous experience, that it would be better to avoid this novel instance of a tiger.

Scene perception adds another layer of abstraction to the visual processing stream. Objects are usually not perceived in isolation, but as elements of a richer context, including other related objects and stimuli. According to Bar (2004), humans learn to predict the objects and their locations typically occurring in a scene based on so-called context frames. Context frames can be viewed as prototypical representations of unique contexts and provide a set of expectations that guide perception, action and eye movements for information gathering. A common empirical finding that supports the context frame idea is that when one is in a familiar context, such as a kitchen, it is easier to recognise objects that are typical of that context, (e.g. a fork) than object that are not (e.g. a bicycle)(Palmer, 1975). In addition to perceptual facilitation, context frames engage a set of context-appropriate responses and stimulus-response rules. For example, in a kitchen the prior tendency to make a coffee is higher than to do a backflip. In cognitive control research, the context-dependent binding of stimuli, actions and outcomes is typically referred to as tasks set. A large body of research provides behavioural (Monsell, 2003) and neural evidence (Dosenbach et al., 2006) for the existence of tasks sets, and that task sets generalise to novel situations (Collins & Frank, 2013). However, the relationship of tasks sets with the perceptual processes of context inference mentioned above remains poorly understood.

Extending the concept of context frames to include a set of metacognitive expectations might be a fruitful approach to explain the results of Study 2. Specifically, inferring a particular context would trigger expectations not only about what one is likely to perceive and how one will respond, but also how much effort one will invest to make these responses. Thus, in some contexts, deliberative and cautious decision making may be appropriate (e.g. when writing a difficult section of a paper), while in another context, a less focused and 'casual' way of decision making may be appropriate (e.g. in a bar with friends). Algorithmically, context frames could adjust the height of a decision threshold a priori to the expected demand for information gathering in a given context. This resonates with the results in Fig 3.3 C, showing that the increase of response time with trial-level conflict was modulated by context. Another interesting question is whether metacognitive expectations (such as the demand for planning) can only become bound to pre-existing abstract context representations, or whether novel abstract control-specific contexts can be formed based on control-relevant properties. Results of Study 2 provide evidence for the latter, suggesting that participants partitioned the state space based on the demand to plan ahead. Participants grouped together the extreme offers 1 and 4 for which a simple strategy was often sufficient and the intermediate offers 2 and 3 for which planning was beneficial. However, further research is needed to unveil the exact nature of control context representations, how they can be learned and how they shape information processing within a larger network of cortical and subcortical brain structures.

So far I discussed that context frames (or specifically control contexts) might be crucial to adjust speed-accuracy trade-offs of decision making to the demands of the current context. However, it is an open question how these context frames are activated in the first place. Interestingly, research on visual perception suggests that context frames can be triggered rapidly by coarse global scene information (Bar, 2004). On the neural level, low spatial frequencies of an image might be projected from early visual areas to association areas in the medial temporal and prefrontal cortex, which then endow the inferior temporal cortex with sensory expectations that help to disambiguate incoming high frequency information. This process of context frame activation resonates to some degree with a proposal from the cognitive control literature that task sets can be readily activated by particular stimulus features (Abrahamse et al., 2016). To my knowledge, however, it has not yet been empirically tested whether task sets can be activated by fast low-frequency sensory pathways. In addition, a question more directly related to the work of this dissertation is, whether such fast pathways can also activate control contexts. These questions could be answered by future studies testing how cognitive performance changes when varying the spatial frequency spectrum or presentation time of stimuli for which scene-control associations have been learned.

Abstractions do not only occur in the domain of states, e.g. as manifested in the perception of scenes, but also in the domain of actions. For example, the abstract action of making a coffee involves a series of intermediate actions, such as picking up the coffee container, pouring in the coffee powder and turning on the coffee maker, with intermediate actions themselves consisting of more primitive motor actions. On the computational level, (temporal) action abstractions have been formalized in hierarchical reinforcement learning (HRL; M. M. Botvinick et al., 2009; Sutton & Barto, 2018). HRL involves the construction of a set of high-level actions that chunk together sequences of more primitive actions. Learning or planning within this abstract action space, thus allows an agent to efficiently select and execute high-level actions with large excepted rewards. An important question here is how action abstractions are discovered in the first place. While the most basic form of action chunking might depend on innate neurophysiological priors, such as with the central pattern generators for walking control (Marder & Bucher, 2001), there is also evidence that state abstractions can emerge from structural learning. It has been shown that humans discover hierarchical representations that allow for efficient planning and appropriate behaviour across a set of possible future tasks (Solway et al., 2014; Tomov et al., 2020).

Konidaris (2019) notes that understanding the creation of useful representations requires to integrate both state abstraction and action abstraction. The state or perceptual abstractions we make might depend on the set of available action abstractions and the other way round, action abstraction might depend on the set of available state abstractions. An interesting observation in this regard is that when perceptual entities become associated with actions, their degrees of abstraction seem to be matched. For example inferring that one stands in front of a coffee machine might engage response tendencies to press the power button, while inferring to be in a kitchen context might engage the more abstract action, coding for the entire sequence of making a coffee. The work in this dissertation suggests that the utility of a representation for the regulations of forward planning, or more generally, for meta-control, may be yet another key to understanding the creation of abstract representations.

4.3 Limitations and future directions

In the current work, forward planning was modelled as solving a finite-horizon MDP using dynamic programming (see equations 3.4-3.6), a method typically employed in model-based reinforcement-learning. Using dynamic programming the optimal policy and state-action values were computed, which could be equivalently obtained by exhaustive forward planning. To account for participants' limited processing capacities, responses were modelled by feeding optimal state-action values into a sigmoid function, including parameters for decision noise and response bias or preference (see sections 2.4.5 and 3.3.4 for further parameters and details). In our analysis, the preference parameters were used to characterise the participants' use of a heuristic or simple decision strategy over a planning strategy. However, a limitation of this interpretation is that the extent to which a heuristic generation process is attributed to a response depends on the specific implementation of the planning process. Thus, in some situations for which we inferred a heuristic strategy, participants might actually have planned ahead, but in a different way than our planning model. Previous studies suggest that humans can adapt their planning in various ways, for example by evaluating only a subset of actions sequences (Huys et al., 2012) or by limiting the depth of planning (Juechems et al., 2019; Keramati et al., 2016). Dissociating those different algorithmic implementations of the planning process based on choice outcomes alone is inherently challenging and probably requires a different experimental design than used in this work. Future work should draw upon richer data (and modelling, see below), including reaction times and neural signals to further pin down the planning processes in the human brain.

In study 2 we used reaction times as an indicator for the duration of participants' planning and analysed its relationship with value-conflict using a linear regression model. A limitation of this approach is that it does not directly link the process of planning with the generation of response times. A fruitful approach for future studies could be to model planning as the refinement of a priori value estimates by the selective sampling of forward sequences that continues until a sufficient amount of evidence is available for one of the considered choice options. On the behavioural level such a model would allow for predictions of reaction times and response outcomes. On the neural level it might allow for novel predictions of how multiple brain networks for planning, valuation and control interact during decision making. For example, it could be tested how sampled forward sequences from the model relate to sequential activity in the hippocampus (Bakkour et al., 2019; Doll et al., 2015; Johnson & Redish, 2007; Redish, 2016), how information obtained by this sampling influences value signals in the vmPFC (Levy & Glimcher, 2012), and how the dACC regulates the planning process (B. Schmidt et al., 2019) by monitoring dynamic value signals and their uncertainties. Recent cognitive models begin to integrate multiple aspects of the decision process, for example by replacing the choice function of reinforcement learning models with a sequential sampling model (Fontanesi et al., 2019; Pedersen & Frank, 2020) or by modelling information sampling as a function of evolving uncertain value estimates (Callaway et al., 2021; D. G. Lee & Daunizeau, 2021). So far these models have only been applied to simple economic choice. Extending these models to sequential decision making tasks that require forward planning would be an interesting avenue for future research.

The results of this dissertation provided evidence that the regulation of planning (i.e. meta-control) relies at least partially on an abstract representation of the state space (control contexts), however, our implementation of planning itself operates on a fine-grained representation of the state space (assuming complete knowledge of the MDP). While it is possible that the brain hosts separate representations of different granularity for different tasks, there is also evidence that people plan based on abstract states and actions (e.g. Tomov et al., 2020; see also section 4.2). Such hierarchical planning could implicitly reduce computational costs of planning, in addition to the top-down regulatory processes studied in the current work. It is uncertain how our results were influenced by this, but future studies should directly test the cost-cutting effect of different abstraction levels on planning.

Complex sequential tasks like the ones developed in this thesis or in previous studies (e.g. Economides et al., 2014; Juechems et al., 2019; Kolling et al., 2014; Korn & Bach, 2018; Schwartenbeck et al., 2015), might provide novel insights into how humans leverage their structured world-knowledge for planning and decision making. Still, challenges remain, as the additional complexity of the tasks permits a multitude of possible hypotheses (see discussion above). Distinguishing between these hypotheses requires most likely rich behavioural and neural measurements, as well as careful computational modelling that bridges the gaps between behaviour, cognitive processes and brain activity. Ultimately, the approach of using rich naturalistic sequential tasks combined with cognitive computational modelling, may help us to better understand why humans in the real world often do not sufficiently consider the future consequences of their actions. Emblematic examples of this human limitation are the overexploitation of limited natural resources or the anthropogenic climate change. Insufficient forward planning, i.e. the mental simulation of action consequences, may indeed be an important contributory cause of these societal problems.

5 References

Abrahamse, E., Braem, S., Notebaert, W., & Verguts, T. (2016). Grounding cognitive control in associative learning. *Psychological bulletin, 142*(7), 693.

Andersen, R. A., & Cui, H. (2009). Intention, action planning, and decision making in parietal-frontal circuits. *Neuron, 63*(5), 568-583.

Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *Neuroimage, 26*(3), 839-851.

Bakkour, A., Palombo, D. J., Zylberberg, A., Kang, Y. H., Reid, A., Verfaellie, M., . . . Shohamy, D. (2019). The hippocampus supports deliberation during value-based decisions. *eLife, 8*, e46080.

Ballard, T., Yeo, G., Neal, A., & Farrell, S. (2016). Departures from optimality when pursuing multiple approach or avoidance goals. *Journal of Applied Psychology, 101*(7), 1056.

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience, 5*(8), 617-629.

Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage, 76*, 412-427.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Behrens, T. E., Muller, T. H., Whittington, J. C., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron, 100*(2), 490-509.

Bejjani, C., Zhang, Z., & Egner, T. (2018). Control by association: Transfer of implicitly primed attentional states across linked stimuli. *Psychonomic Bulletin & Review, 25*(2), 617-626.

Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., & West, M. (2003). *Non-centered parameterisations for hierarchical models and data augmentation.* Paper presented at the Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting.

Bhandari, A., Badre, D., & Frank, M. J. (2017). Learning cognitive control. *The Wiley handbook of cognitive control. John Wiley & Sons.*

Biderman, N., Bakkour, A., & Shohamy, D. (2020). What are memories for? The hippocampus bridges past experience with future decisions. *Trends in Cognitive Sciences, 24*(7), 542-556.

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., . . . Goodman, N. D. (2018). Pyro: Deep universal probabilistic programming. *arXiv preprint arXiv:1810.09538*.

Blair, K., Marsh, A. A., Morton, J., Vythilingam, M., Jones, M., Mondillo, K., . . . Blair, J. R. (2006). Choosing the lesser of two evils, the better of two goods: specifying the roles of ventromedial prefrontal cortex and dorsal anterior cingulate in object choice. *Journal of Neuroscience, 26*(44), 11379-11386.

Bossaerts, P. (2010). Risk and risk prediction error signals in anterior insula. *Brain Structure and Function, 214*(5-6), 645-653.

Botvinick, M., & Braver, T. (2015). Motivation and cognitive control: from behavior to neural mechanism. *Annual Review of Psychology, 66*, 83-113.

Botvinick, M., Wang, J. X., Dabney, W., Miller, K. J., & Kurth-Nelson, Z. (2020). Deep reinforcement learning and its neuroscientific implications. *Neuron*.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological review, 108*(3), 624.

Botvinick, M. M., & Cohen, J. D. (2014). The computational and neural basis of cognitive control: charted territory and new frontiers. *Cognitive science, 38*(6), 1249-1285.

Botvinick, M. M., Niv, Y., & Barto, A. G. (2009). Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition, 113*(3), 262-280.

Boureau, Y.-L., Sokol-Hessner, P., & Daw, N. D. (2015). Deciding how to decide: self-control and meta-decision making. *Trends in Cognitive Sciences, 19*(11), 700-710.

Braem, S., Verguts, T., Roggeman, C., & Notebaert, W. (2012). Reward modulates adaptations to conflict. *Cognition, 125*(2), 324-332.

Bugg, J. M., & Dey, A. (2018). When stimulus-driven control settings compete: On the dominance of categories as cues for control. *Journal of experimental psychology: human perception and performance, 44*(12), 1905.

Bugg, J. M., Jacoby, L. L., & Chanani, S. (2011). Why it is too early to lose control in accounts of item-specific proportion congruency effects. *Journal of experimental psychology: human perception and performance, 37*(3), 844.

Busemeyer, J. R., Gluth, S., Rieskamp, J., & Turner, B. M. (2019). Cognitive and Neural Bases of Multi-Attribute, Multi-Alternative, Value-based Decisions. *Trends in Cognitive Sciences*.

Bustamante, L., Lieder, F., Musslick, S., Shenhav, A., & Cohen, J. (2021). Learning to overexert cognitive control in a Stroop task. *Cognitive, Affective, & Behavioral Neuroscience, 21*(3), 453-471.

Callaway, F., Rangel, A., & Griffiths, T. L. (2021). Fixation patterns in simple choice reflect optimal information sampling. *PLoS computational biology, 17*(3), e1008863.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., . . . Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76*(1).

Cavanagh, J. F., Wiecki, T. V., Cohen, M. X., Figueroa, C. M., Samanta, J., Sherman, S. J., & Frank, M. J. (2011). Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. *Nature neuroscience, 14*(11), 1462-1467.

Chiu, Y.-C., & Egner, T. (2019). Cortical and subcortical contributions to context-control learning. *Neuroscience & Biobehavioral Reviews, 99*, 33-41.

Collins, A. G., & Cockburn, J. (2020). Beyond dichotomies in reinforcement learning. *Nature Reviews Neuroscience, 21*(10), 576-586.

Collins, A. G., & Frank, M. J. (2013). Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychological review, 120*(1), 190.

Craig, A. D., & Craig, A. (2009). How do you feel--now? The anterior insula and human awareness. *Nature Reviews Neuroscience, 10*(1).

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron, 69*(6), 1204-1215.

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature neuroscience, 16*(1), 105.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron, 73*(3), 415-434.

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron, 80*(2), 312-325.

Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D., & Daw, N. D. (2015). Model-based choices involve prospective neural activity. *Nature neuroscience, 18*(5), 767.

Dosenbach, N. U., Visscher, K. M., Palmer, E. D., Miezin, F. M., Wenger, K. K., Kang, H. C., . . . Petersen, S. E. (2006). A core system for the implementation of task sets. *Neuron, 50*(5), 799-812.

Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. *Trends in Cognitive Sciences, 14*(4), 172-179.

Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in neurosciences, 23*(10), 475-483.

Ebitz, R. B., & Hayden, B. Y. (2016). Dorsal anterior cingulate: a Rorschach test for cognitive neuroscience. *Nature neuroscience, 19*(10), 1278.

Economides, M., Guitart-Masip, M., Kurth-Nelson, Z., & Dolan, R. J. (2014). Anterior cingulate cortex instigates adaptive switches in choice by integrating immediate and delayed components of value in ventromedial prefrontal cortex. *Journal of Neuroscience, 34*(9), 3340-3349.

Economides, M., Guitart-Masip, M., Kurth-Nelson, Z., & Dolan, R. J. (2015). Arbitration between controlled and impulsive choices. *Neuroimage, 109*, 206-216.

Egner, T. (2014). Creatures of habit (and control): a multi-level learning perspective on the modulation of congruency effects. *Frontiers in psychology, 5*, 1247.

Eppinger, B., Goschke, T., & Musslick, S. (2021). Meta-control: From psychology to computational neuroscience. *Cognitive, Affective, & Behavioral Neuroscience*, 1-6.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & psychophysics, 16*(1), 143-149.

Fontanesi, L., Gluth, S., Spektor, M. S., & Rieskamp, J. (2019). A reinforcement learning diffusion decision model for value-based decisions. *Psychonomic Bulletin & Review, 26*(4), 1099-1121.

Frank, M. J. (2006). Hold your horses: a dynamic computational role for the subthalamic nucleus in decision making. *Neural networks, 19*(8), 1120-1136.

Frank, M. J., Gagne, C., Nyhus, E., Masters, S., Wiecki, T. V., Cavanagh, J. F., & Badre, D. (2015). fMRI and EEG predictors of dynamic decision parameters during human reinforcement learning. *Journal of Neuroscience, 35*(2), 485-494.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127-138.

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical transactions of the Royal Society B: Biological sciences, 364*(1521), 1211-1221.

Fuster, J. (2015). *The prefrontal cortex*: Academic Press.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, 733-760.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science, 7*(4), 457-472.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science, 349*(6245), 273-278.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology, 62*, 451-482.

Gluth, S., Rieskamp, J., & Büchel, C. (2012). Deciding when to decide: time-variant sequential sampling models explain the emergence of value-based decisions in the human brain. *Journal of Neuroscience, 32*(31), 10686-10698.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual review of neuroscience, 30*.

Goschke, T. (2013). Volition in action: intentions, control dilemmas and the dynamic regulation of intentional control. *Action science: Foundations of an emerging discipline*, 409-434.

Goschke, T. (2014). Dysfunctions of decision-making and cognitive control as transdiagnostic mechanisms of mental disorders: advances, gaps, and needs in current research. *International journal of methods in psychiatric research, 23*(S1), 41-57.

Gratton, G., Coles, M. G., & Donchin, E. (1992). Optimizing the use of information: strategic control of activation of responses. *Journal of Experimental Psychology: General, 121*(4), 480.

Griffiths, T. L., Callaway, F., Chang, M. B., Grant, E., Krueger, P. M., & Lieder, F. (2019). Doing more with less: meta-reasoning and meta-learning in humans and machines. *Current Opinion in Behavioral Sciences, 29*, 24-30.

Hare, T. A., Schultz, W., Camerer, C. F., O'Doherty, J. P., & Rangel, A. (2011). Transformation of stimulus value signals into motor commands during simple choice. *Proceedings of the National Academy of Sciences, 108*(44), 18120-18125.

Hayes-Roth, B., & Hayes-Roth, F. (1979). A cognitive model of planning. *Cognitive science, 3*(4), 275-310.

Heilbronner, S. R., & Hayden, B. Y. (2016). Dorsal anterior cingulate cortex: a bottom-up view. *Annual review of neuroscience, 39*, 149-170.

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research, 14*(1), 1303-1347.

Hunt, L., Daw, N., Kaanders, P., MacIver, M., Mugan, U., Procyk, E., . . . Stachenfeld, K. (2021). Formalizing planning and information search in naturalistic decision-making. *Nature neuroscience*, 1-14.

Hunt, L. T., Kolling, N., Soltani, A., Woolrich, M. W., Rushworth, M. F., & Behrens, T. E. (2012). Mechanisms underlying cortical activity during value-guided choice. *Nature neuroscience, 15*(3), 470-476.

Hutton, C., Bork, A., Josephs, O., Deichmann, R., Ashburner, J., & Turner, R. (2002). Image distortion correction in fMRI: a quantitative evaluation. *Neuroimage, 16*(1), 217-240.

Huys, Q. J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: how the Pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS computational biology, 8*(3), e1002410.

Huys, Q. J., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., Gershman, S. J., . . . Roiser, J. P. (2015). Interplay of approximate planning strategies. *Proceedings of the National Academy of Sciences, 112*(10), 3098-3103.

Jahfari, S., Waldorp, L., van den Wildenberg, W. P., Scholte, H. S., Ridderinkhof, K. R., & Forstmann, B. U. (2011). Effective connectivity reveals important roles for both the hyperdirect (fronto-subthalamic) and the indirect (fronto-striatal-pallidal) fronto-basal ganglia pathways during response inhibition. *Journal of Neuroscience, 31*(18), 6891-6899.

Johnson, A., & Redish, A. D. (2007). Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *Journal of Neuroscience, 27*(45), 12176-12189.

Juechems, K., Balaguer, J., Castañón, S. H., Ruz, M., O’Reilly, J. X., & Summerfield, C. (2019). A network for computing value equilibrium in the human medial prefrontal cortex. *Neuron, 101*(5), 977-987. e973.

Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS computational biology, 7*(5), e1002055.

Keramati, M., Smittenaar, P., Dolan, R. J., & Dayan, P. (2016). Adaptive integration of habits into depth-limited planning defines a habitual-goal–directed spectrum. *Proceedings of the National Academy of Sciences, 113*(45), 12868-12873.

Kerns, J. G., Cohen, J. D., MacDonald, A. W., Cho, R. Y., Stenger, V. A., & Carter, C. S. (2004). Anterior cingulate conflict monitoring and adjustments in control. *Science, 303*(5660), 1023-1026.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception, 36*(14), 1.

Kolling, N., Scholl, J., Chekroud, A., Trier, H. A., & Rushworth, M. F. (2018). Prospection, perseverance, and insight in sequential behavior. *Neuron, 99*(5), 1069-1082. e1067.

Kolling, N., Wittmann, M., & Rushworth, M. F. (2014). Multiple neural mechanisms of decision making and their competition under changing risk pressure. *Neuron, 81*(5), 1190-1202.

Konidaris, G. (2019). On the necessity of abstraction. *Current Opinion in Behavioral Sciences, 29*, 1-7.

Kool, W., Gershman, S. J., & Cushman, F. A. (2017). Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychological science, 28*(9), 1321-1333.

Kool, W., Gershman, S. J., & Cushman, F. A. (2018). Planning complexity registers as a cost in metacontrol. *Journal of cognitive neuroscience, 30*(10), 1391-1404.

Korn, C. W., & Bach, D. R. (2018). Heuristic and optimal policy computations in the human brain during sequential decision-making. *Nature communications, 9*(1), 325.

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature neuroscience, 13*(10), 1292-1298.

Krebs, R. M., Boehler, C. N., Appelbaum, L. G., & Woldorff, M. G. (2013). Reward associations reduce behavioral interference by changing the temporal dynamics of conflict processing. *PloS one, 8*(1), e53894.

Krebs, R. M., Boehler, C. N., & Woldorff, M. G. (2010). The influence of reward associations on conflict processing in the Stroop task. *Cognition, 117*(3), 341-347.

Kumar, R., Carroll, C., Hartikainen, A., & Martín, O. A. (2019). ArviZ a unified library for exploratory analysis of Bayesian models in Python.

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *Behavioral and brain sciences, 36*(6), 661-679.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software, 82*(13).

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences, 40*.

Lee, D. G., & Daunizeau, J. (2021). Trading mental effort for confidence in the metacognitive control of value-based decision-making. *eLife, 10*, e63282.

Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron, 81*(3), 687-699.

Levy, D. J., & Glimcher, P. W. (2012). The root of all value: a neural common currency for choice. *Current opinion in neurobiology, 22*(6), 1027-1038.

Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological review, 124*(6), 762.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences, 43*.

Lieder, F., Shenhav, A., Musslick, S., & Griffiths, T. L. (2018). Rational metareasoning and the plasticity of cognitive control. *PLoS computational biology, 14*(4), e1006043.

Loued-Khenissi, L., Pfeuffer, A., Einhäuser, W., & Preuschoff, K. (2020). Anterior insula reflects surprise in value-based decision-making and perception. *Neuroimage, 210*, 116549.

MacDonald, A. W., Cohen, J. D., Stenger, V. A., & Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science, 288*(5472), 1835-1838.

Maldjian, J. A., Laurienti, P. J., Kraft, R. A., & Burdette, J. H. (2003). An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *Neuroimage, 19*(3), 1233-1239.

Marder, E., & Bucher, D. (2001). Central pattern generators and the control of rhythmic movements. *Current Biology, 11*(23), R986-R996.

Marković, D., Goschke, T., & Kiebel, S. J. (2020). Meta-control of the exploration-exploitation dilemma emerges from probabilistic inference over a hierarchy of time scales. *Cognitive, Affective, & Behavioral Neuroscience*, 1-25.

Martin, A. (2007). The representation of object concepts in the brain. *Annu. Rev. Psychol., 58*, 25-45.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual review of neuroscience, 24*(1), 167-202.

Miller, K. J., & Venditto, S. J. C. (2020). Multi-step Planning in the Brain. *Current Opinion in Behavioral Sciences, 38*, 29-39.

Monosov, I. E., Haber, S. N., Leuthardt, E. C., & Jezzini, A. (2020). Anterior cingulate cortex and the control of dynamic behavior in primates. *Current Biology, 30*(23), R1442-R1454.

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences, 7*(3), 134-140.

Mormann, M. M., Malmaud, J., Huth, A., Koch, C., & Rangel, A. (2010). The drift diffusion model can account for the accuracy and reaction time of value-based choices under high and low time pressure. *Judgment and Decision Making, 5*(6), 437-449.

Neal, A., Ballard, T., & Vancouver, J. B. (2017). Dynamic Self-Regulation and Multiple-Goal Pursuit. *Annual Review of Organizational Psychology and Organizational Behavior, 4*, 401-423.

Niendam, T. A., Laird, A. R., Ray, K. L., Dean, Y. M., Glahn, D. C., & Carter, C. S. (2012). Meta-analytic evidence for a superordinate cognitive control network subserving diverse executive functions. *Cognitive, Affective, & Behavioral Neuroscience, 12*(2), 241-268.

Orehek, E., & Vazeou-Nieuwenhuis, A. (2013). Sequential and concurrent strategies of multiple goal pursuit. *Review of General Psychology, 17*(3), 339.

Ostwald, D., Bruckner, R., & Heekeren, H. (2018). *Computational mechanisms of human state-action-reward contingency learning under perceptual uncertainty*. Paper presented at the Conference on Cognitive Computational Neuroscience, Philadelphia, Pennsylvania, USA.

Ott, F., Marković, D., Strobel, A., & Kiebel, S. J. (2020). Dynamic integration of forward planning and heuristic preferences during multiple goal pursuit. *PLoS computational biology, 16*(2), e1007685.

Padmala, S., & Pessoa, L. (2011). Reward reduces conflict by enhancing attentional control and biasing visual cortical processing. *Journal of cognitive neuroscience, 23*(11), 3419-3432.

Palmer, t. E. (1975). The effects of contextual scenes on the identification of objects. *Memory & cognition, 3*, 519-526.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., . . . Lerer, A. (2017). Automatic differentiation in pytorch.

Pedersen, M. L., & Frank, M. J. (2020). Simultaneous hierarchical bayesian parameter estimation for reinforcement learning and drift diffusion models: a tutorial and links to neural data. *Computational Brain & Behavior, 3*, 458-471.

Piray, P., & Daw, N. D. (2021). Linear reinforcement learning in planning, grid fields, and cognitive control. *Nature communications, 12*(1), 1-20.

Pochon, J.-B., Riis, J., Sanfey, A. G., Nystrom, L. E., & Cohen, J. D. (2008). Functional imaging of decision conflict. *Journal of Neuroscience, 28*(13), 3468-3473.

Polson, N. G., & Scott, J. G. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian statistics, 9*, 501-538.

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*: John Wiley & Sons.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: current issues and history. *Trends in Cognitive Sciences, 20*(4), 260-281.

Redish, A. D. (2016). Vicarious trial and error. *Nature Reviews Neuroscience, 17*(3), 147-159.

Schacter, D. L., Addis, D. R., Hassabis, D., Martin, V. C., Spreng, R. N., & Szpunar, K. K. (2012). The future of memory: remembering, imagining, and the brain. *Neuron, 76*(4), 677-694.

Schmidt, A. M., & DeShon, R. P. (2007). What to do? The effects of discrepancies, incentives, and time on dynamic goal prioritization. *Journal of Applied Psychology, 92*(4), 928.

Schmidt, A. M., & Dolis, C. M. (2009). Something's got to give: The effects of dual-goal difficulty, goal progress, and expectancies on resource allocation. *Journal of Applied Psychology, 94*(3), 678.

Schmidt, B., Duin, A. A., & Redish, A. D. (2019). Disrupting the medial prefrontal cortex alters hippocampal sequences during deliberative decision making. *Journal of neurophysiology, 121*(6), 1981-2000.

Schwartenbeck, P., FitzGerald, T. H., Mathys, C., Dolan, R., & Friston, K. (2015). The dopaminergic midbrain encodes the expected certainty about desired outcomes. *Cerebral Cortex, 25*(10), 3434-3445.

Schwöbel, S., Marković, D., Smolka, M. N., & Kiebel, S. J. (2021). Balancing control: a Bayesian interpretation of habitual and goal-directed behavior. *Journal of mathematical psychology, 100*, 102472.

Seabold, S., & Perktold, J. (2010). *Statsmodels: Econometric and statistical modeling with python.* Paper presented at the Proceedings of the 9th Python in Science Conference.

Shadlen, M. N., & Shohamy, D. (2016). Decision making and sequential sampling from memory. *Neuron, 90*(5), 927-939.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron, 79*(2), 217-240.

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). Toward a rational and mechanistic account of mental effort. *Annual review of neuroscience, 40*, 99-124.

Shenhav, A., Straccia, M. A., Cohen, J. D., & Botvinick, M. M. (2014). Anterior cingulate engagement in a foraging context reflects choice difficulty, not foraging value. *Nature neuroscience, 17*(9), 1249-1254.

Simon, D. A., & Daw, N. D. (2011). Neural correlates of forward planning in a spatial decision task in humans. *Journal of Neuroscience, 31*(14), 5526-5539.

Simon, J. R. (1969). Reactions toward the source of stimulation. *Journal of experimental psychology, 81*(1), 174.

Singer, T., Critchley, H. D., & Preuschoff, K. (2009). A common role of insula in feelings, empathy and uncertainty. *Trends in Cognitive Sciences, 13*(8), 334-340.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts* (Vol. 9): Harvard University Press Cambridge, MA.

Smith, E. H., Horga, G., Yates, M. J., Mikell, C. B., Banks, G. P., Pathak, Y. J., . . . Botvinick, M. M. (2019). Widespread temporal coding of cognitive control in the human prefrontal cortex. *Nature neuroscience*, 1-9.

Soltani, A., Khorsand, P., Guo, C., Farashahi, S., & Liu, J. (2016). Neural substrates of cognitive biases during probabilistic inference. *Nature communications, 7*, 11393.

Solway, A., Diuk, C., Córdova, N., Yee, D., Barto, A. G., Niv, Y., & Botvinick, M. M. (2014). Optimal behavioral hierarchy. *PLoS computational biology, 10*(8), e1003779.

Stan Development Team. (2018). *PyStan: the Python interface to Sta*n, Version 2.19.1.1. http://mc-stan.org.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of experimental psychology, 18*(6), 643.

Stürmer, B. (2011). Reward and punishment effects on error processing and conflict control. *Frontiers in psychology, 2*, 335.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*: MIT press.

Tajima, S., Drugowitsch, J., & Pouget, A. (2016). Optimal policy for value-based decision-making. *Nature communications, 7*(1), 1-12.

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., & Gelman, A. (2018). Validating Bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*.

Team, R. C. (2013). R: A language and environment for statistical computing.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science, 331*(6022), 1279-1285.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological review, 55*(4), 189.

Tomov, M. S., Yagati, S., Kumar, A., Yang, W., & Gershman, S. J. (2020). Discovery of hierarchical representations for efficient planning. *PLoS computational biology, 16*(4), e1007594.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*(5), 1413-1432.

Venkatraman, V., Rosati, A. G., Taren, A. A., & Huettel, S. A. (2009). Resolving response, decision, and strategic control: evidence for a functional topography in dorsomedial prefrontal cortex. *Journal of Neuroscience, 29*(42), 13158-13164.

Wang, S., Feng, S. F., & Bornstein, A. (2020). Mixing memory and desire: How memory reactivation supports deliberative decision-making.

White, J. K., Bromberg-Martin, E. S., Heilbronner, S. R., Zhang, K., Pai, J., Haber, S. N., & Monosov, I. E. (2019). A neural network for information seeking. *Nature communications, 10*(1), 1-19.

Wiecki, T. V., & Frank, M. J. (2013). A computational model of inhibitory control in frontal cortex and basal ganglia. *Psychological review, 120*(2), 329.

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in Python. *Frontiers in neuroinformatics, 7*, 14.

Yee, E. (2019). Abstraction and concepts: when, how, where, what and why? In: Taylor & Francis.

Declaration according to § 5 of the doctoral regulations

Assurance

I hereby certify that I have authored this thesis without the undue assistance of third parties and without the use of aids other than those indicated; the assistance of third parties has only been used to an extent that is scientifically justifiable and permissible under examination law; the ideas taken directly or indirectly from external sources are identified as such. No inadmissible pecuniary benefits, either direct or indirect, have been paid to third parties in connection with the content of this dissertation. The thesis has not been submitted to any other examination authority in the same or a similar form, either in Germany or abroad.

The submitted dissertation was written at the Institute of General Psychology, Biopsychology and Methods of Psychology at the chair of Neuroimaging at Technische Universität Dresden under the supervision of Prof. Dr. Stefan Kiebel and Prof. Dr. Alexander Strobel.

No previous unsuccessful doctoral examination procedures have taken place.

The doctoral regulations of the Department of Mathematics and Natural Sciences of Technische Universität Dresden, in the version of 23.02.2011, last amendment 23.05.2018, are acknowledged.