Joint modeling of choices and reaction times based on Bayesian contextual behavioral control

Sarah Schwöbel[1*], **Dimitrije Markovic**[1], **Michael N. Smolka**[2], **Stefan Kiebel**[1,3]

***For correspondence:**
sarah.schwoebel@tu-dresden.de

[1]Department of Psychology, Technische Universität Dresden, Dresden, Germany; [2]Department of Psychiatry and Neuroimaging Center, Technische Universität Dresden, Dresden, Germany; [3]Centre for Tactile Internet with Human-in-the-Loop (CeTI), Technische Universität Dresden, Dresden, Germany

Abstract In cognitive neuroscience and psychology, reaction times are an important behavioral measure. However, in instrumental learning and goal-directed decision making experiments, findings often rely only on choice probabilities from a value-based model, instead of reaction times. Recent advancements have shown that it is possible to connect value-based decision models with reaction time models, for example in a joint reinforcement learning and diffusion decision model. We propose a novel joint model of both choices and reaction times by combining a mechanistic account of Bayesian sequential decision making with a sampling procedure. Specifically, we use a recent context-specific Bayesian forward planning model which we extend by an MCMC sampler to obtain both choices and reaction times. We show that we can explain and reproduce a rather wide range of well-known experimental findings in value based-decision making as well as classical inhibition and switching tasks. First, we use the proposed model to explain how instrumental learning and automatized behavior result in decreased reaction times and improved accuracy. Second, we reproduce classical results in the Eriksen flanker task. Third, we reproduce established findings in task switching. These findings show that the proposed joint behavioral model may describe common underlying processes in all these types of decision making paradigms.

Introduction

Many key findings in psychology and cognitive neuroscience of the last decades are based on the measurement and analysis of both response accuracy and reaction times in behavioral experiments. For example, changes in both mean reaction times and response accuracy during and after conflicting decisions are typically interpreted to demonstrate underlying decision making processes. Such effects of classical experimental paradigms are remarkably stable and have also been used to show how decision making is impaired in several mental disorders (***Goschke, 2014***; ***Gratton et al., 2018***; ***Kozak and Cuthbert, 2016***; ***Gratton et al., 1992***; ***Stins et al., 2007***; ***Kiesel et al., 2010***; ***Monsell, 2003***).

Two important aspects that many of these classical experiments typically do not take into account are the influence of uncertainty and reward structure. To study these aspects, often sequential decision making tasks are used in the instrumental learning and value-based decision making literature, e.g. (***Daw et al., 2011***; ***Kolling et al., 2014***). Here, the experimental findings usually do not rely on reaction time effects, but on a value-based decision model and associated choice prob-

abilities. These allow to infer model parameters that are used to characterize the underlying mechanism and explain inter-individual differences. The value-based decision making models typically used for such experiments are either variants of reinforcement learning models or Bayesian (active inference) models. However, in their original variants, these models do not describe reaction times associated with decisions.

Recently, there have been some successful joint applications of value-based decision models, specifically reinforcement learning, and reaction time models (***Milosavljevic et al., 2010***; ***Pedersen et al., 2017***; ***Fontanesi et al., 2019***; ***Shahar et al., 2019***; ***Miletić et al., 2021***), which are typically evidence accumulator models such as diffusion decision models (DDM) (***Ratcliff, 1978***; ***Forstmann and Wagenmakers, 2015***) and so-called race diffusion models (***Milosavljevic et al., 2010***). Here, the principled idea is to connect choice values and probabilities to reaction times by linking parameters in both models. For example, the trial-wise expected reward (Q-values) in reinforcement learning models can be used to vary the drift rate of a DDM (***Pedersen et al., 2017***). This approach has recently been extended to multi-choice tasks using race diffusion models, where instead of having one accumulator as in the DDM, each available choice option is associated with a different accumulator (***Fontanesi et al., 2019***; ***Miletić et al., 2021***).

In this article, we will build upon and extend these recent approaches in combining value-based models with reaction times. We propose a novel model that accounts for three key components of behavioural effects: (i) repetition of responses, (ii) context-specific effects, and (iii) uncertainty at multiple levels. We will show that the combination of these three model components allows us to model a surprisingly wide range of findings, from classical reaction-time based cognitive control tasks to sequential value-based decision tasks.

As a value-based decision model we use a recently introduced Bayesian prior-based contextual control model for forward planning and goal-directed decision making (***Schwöbel et al., 2021***). This model has two key ingredients: (i) a prior over actions or action sequences which is learned based on repetitions and independent of reward expectations, and (ii) differences in the environment are described as different contexts that need to be inferred. To link internal variables of this model to choices and reaction times, we used, in a straight-forward fashion, an independent Markov chain Monte Carlo (MCMC) sampler. The resulting approach has the advantage over previous approaches that it allows us to explain reaction times mechanistically and specifically describes the actual functional role of sampling and noise, and how this noise consequently shapes reaction time distributions. In other words, in the proposed model, the sampling process is not simply a mapping from key variables of the underlying value-based model to choice and reaction time distributions as in DDM-based models but an integral part of the decision making machinery.

To illustrate this novel approach of jointly modeling choices and reaction times, we present three simulations that capture several well-established experimental findings. We use a sequential value-based decision task and two well-known cognitive control tasks, the Eriksen flanker task and task switching, which are commonly used as experimental paradigms to study cognitive processes and impaired decision making in mental disorders (***Goschke, 2014***; ***Gratton et al., 2018***). One overarching result was that behavioural phenomena of these three rather different tasks were replicated by the same model using only small and well-motivated changes in task-dependent parameterizations. Specifically, first, using the value-based decision task, we show that our approach explains decreasing reaction times when learning goal-directed actions. As an agent becomes more certain about its context, reaction times decrease. Additionally, if a prior over behavior is learned, it facilitates faster and more reliable behavior, even up to the point that automatic, goal-independent behavior emerges. Second, we demonstrate that the novel approach can in principle also be generalized to experiments and phenomena which would usually not be labelled as value-based decision making. To show this, we provide a mechanistic understanding of experimental phenomena in the Flanker task, a typical cognitive control task. This task is often interpreted to measure interference control which is a component of inhibitory control and we show how the conflict can be re-interpreted as a conflict between priors and contexts. Third, we demonstrate

that we can provide an explanation for choices and reaction times in task switching, which is typically interpreted to measure updating of task-relevant information. We show that the effects can be interpreted as updating of context-specific representations as participants switch between two contexts. We close by discussing the implications of the underlying mechanism and its relation to alternative models.

Methods

In this section, we sketch the main idea behind the computational model of behavior used for simulations, and the novel reaction time algorithm which was used to generate agent's reaction times.

Prior-based contextual control model

The computational model of behavior is based on a hierarchical partially observable Markov decision process which is represented as a Bayesian generative model. In this model, the dynamics of the environment are clustered into episodes of fixed length, where the transition and reward dynamics in an episode are determined by the current context. Meaning, an agent represents differences in the environment as separate Markov decision processes (MDPs) which are learned for each context separately. Following ideas of active inference (***Friston et al., 2015***, ***2016***) and the free energy principle (***Friston, 2009***, ***2010***), an agent inverts this generative model using approximate inference based on the variational free energy. More specifically, the computational model is an extension of the recently proposed hierarchical Bayesian forward planning and habit learning model (***Schwöbel et al., 2021***). Using this approach, we enable an agent to infer beliefs about future states, rewards, the context, as well as probabilities of choosing different actions or sequences of actions. In accordance with the active inference literature (***Friston et al., 2015***, ***2016***) we will call a deterministic sequence of actions a policy.

Concretely, in each episode, the agent infers the active context and loads the corresponding MDP. Using this MDP, it infers future states and rewards, and based on that posterior probabilities of choosing actions or policies. At the end of an episode, the agent also uses the newly observed rewards to update the parameters of the generative model, therewith learning the structure of the environment.

We will only outline the key ideas of the computational model here, the interested reader can find the full mathematical details in the Supplementary files, as well as its python implementation on github[1]. The main idea can be sketched in a Bayesian equation as

$$p\left(\pi|c,\boldsymbol{R}\right) \propto p\left(\pi|c\right) p\left(\boldsymbol{R}|\pi,c\right) \tag{1}$$

which describes the posterior probability $p\left(\pi|c,\boldsymbol{R}\right)$ of whether an agent should choose policy π, given desired rewards $\boldsymbol{R}$ in the current context c. This posterior is a categorical distribution and according to Bayes' rule it is proportional to the likelihood of obtaining rewards under a policy in the current context $p\left(\boldsymbol{R}|\pi,c\right)$ times a context-specific prior over policies $p\left(\pi|c\right)$.

Importantly, the likelihood of rewards is calculated based on the MDP of the current context, which contains action-outcome contingencies and therewith encodes a goal-directed value of a policy. The prior over policies on the other hand, is updated and learned based on Bayesian learning rules, which yield higher a priori probabilities for policies which have been previously chosen in this context. In our past work (***Schwöbel et al., 2021***) we proposed to interpret this as repetition-based habit learning, as this term implements a priori tendencies to repeat policies, independent of any reward expectations. Due to the prior being over policies, i.e. full action sequences, habits in this model can be additionally interpreted as context-dependent automatized action sequences.

Response conflicts can emerge when the context is not directly observable, and there is uncertainty over the current context that cannot be fully resolved. Here, the agent may not not know

[1] https://github.com/SSchwoebel/BalancingControl

with certainty which rules of the environment currently apply. To enable context inference, we introduced context observations, i.e. cues, into the model, where the agent maintains a generative model of observation probabilities $p\left(o_c|c\right)$. Using Bayesian inversion, as well as the observed events of the context-dependent MDP, and context change probabilities $p\left(c|c'\right)$ the agent can infer a posterior over contexts $p\left(c|o_c\right)$. The resulting posterior over policies

$$p\left(\pi|R\right) \propto \sum_c p\left(\pi|c,R\right) p\left(c|o_c\right) \quad (2)$$

$$= p\left(\pi\right) p\left(R|\pi\right) \quad (3)$$

is then a mix of the context-specific posteriors over policies, weighted by the posterior probabilities of each context. The posterior over policies $p\left(\pi|R\right)$ gives the probability that an agent should choose a specific policy π, given it wants to receive rewards R. To select a policy, an agent will sample from this posterior and execute the corresponding policy, the concrete sampling procedure is described below. This posterior is proportional to the prior over policies $p\left(\pi\right)$ times the likelihood of receiving rewards $p\left(R|\pi\right)$. In a conflict situation, the two conflicting policies would be similarly weighted in the posterior, which leads to an increased error rate, as they would be similarly likely to be selected.

Free parameters and simulation setups

This model has three free parameters which we will vary in the Results section to recreate known reaction time effects in behavioral paradigms (see also Supplementary files): The values of the hyper-parameters of the prior over policies, the expected context transition probability, and the context cue uncertainty. In what follows we will explain the free parameters and in more detail how they connect to experimental setups.

Prior hyper-parameters

As the prior over policies is learned, the parameters of the priors are treated hyper-priors (latent random variables), defined as a Dirichlet distribution. The Dirichlet distribution is parameterised using concentration parameters or pseudo counts α. Essentially, they count how often a policy was chosen in a context, enabling repetition-based learning. While the counting rule is given by the Bayesian updates, the initial values from which counting starts are free parameters which can be chosen at the beginning of a simulation. We defined a so-called habitual tendency $h = \frac{1}{\alpha_{init}}$, where $h = 1.0$ mean the counting starts at 1, giving each new choice a big effect on the prior over policies. Hence we call an agent with high habitual tendency a "strong prior learner". Lower values of h, e.g. $h = 0.001$, means the counting starts at high values, e.g. $\alpha = 1000$, which has the effect that each new choice has little influence on the prior over policies. We call such an agent a "weak prior learner" as in this setting the prior learning is almost neglectable as the pseudo counts are dominated by initial values. In our previous study, we argued that the habitual tendency parameter may be used to model inter-individual differences in habit tasks (***Schwöbel et al., 2021***). We show the effects of strong and weak prior learning on reaction times and accuracy in a sequential decision task (see Section Value-based decision making in a grid world).

However, not all initial pseudo counts need to be set to the same values. In order to model a priori context–response associations, the initial values can be set so that the prior over policies initially has a bias for specific actions or policies in a specific context. Additionally, the pseudo counts can be subjected to a forgetting factor (see Supplementary files) which has them slowly decrease over the course of an experiment, so that later choices still have a measurable effect on the prior over policies. We use such a priori context–response associations to model interference effects in the Flanker task (see Section Flanker task).

Context transition probability

The free parameter of the context transition probability encodes an agent's assumption of how likely a context change is expected to occur after the end of a behavioral episode. For example

in a task switching experiment, two task sets would correspond to two contexts, and the context transition probability encodes how often an agent thinks the current task set will change. If set low, an agents expects to stay within the same context, and even a context cue may not be enough for the agent to infer that the context indeed changed. Traces of the previous context may then still be present even after a context change. If set high, an agent expects a context change to happen after every episode which eases inference of an actual context change, but may also lead to an agent falsely loading the wrong context for action evaluation. We will use this to model inter trial effects in a task switching experiment (see Section Task switching).

Context cue uncertainty

The context cue uncertainty encodes how certain an agent is to have perceived the context cue correctly. For example in a task switching experiment, the current task set is cued, and the uncertainty determines how well an agent perceived the cue. A high uncertainty means an agent may not always rely on the cue and may instead use the previous context to make decisions, while a low uncertainty means an agent perceived the cue well and can reliably load the current context. We use this context cue uncertainty to model known cue presentation time effects on reaction times in a task switching experiment (see Section Task switching).

Task specific adaptations

Note that the general machinery the agent uses for inference and planning is the same in all the setups described above. This means that the process of perceiving a context cue, inferring the context, loading the specific action–outcome contingencies and prior, planning ahead and then sampling an action are common to all setups and all simulations below. For different tasks, we only adapt the agent to the task at hand by changing the three free parameters, which model in a task-dependent fashion (i) how strong the prior is preset or learned, (ii) how well a context cue is perceived and (iii) how stable agents think their environment is. This is well-motivated by the different tasks where, for example, context presentation times also induce uncertainty in cue perception in participants. Importantly, most or all of this machinery is vital in simulating agent behavior in such a wide range of tasks. We also show simulated experiments (in Supplementary files) with parts of the machinery switched off, and show that the typically measured behavioral effects do not arise under these conditions.

Reaction time algorithm

To connect the quantities of the prior-based contextual control model to reaction times and their distributions, we propose here a novel sampling-based reaction time modeling approach that generalizes well to arbitrary numbers of policies.

The idea of this approach is that the prior is learned and stored by the brain, and can quickly be recalled and loaded upon receiving a context cue or inferring a context. This would allow the brain to use the prior as a heuristic based on past choices, which would allow for fast action selection in familiar situations, or when there is a tight deadline for selecting the action. The goal-directed likelihood on the other hand may be slow and costly to evaluate fully, as it is based on an MDP. We propose that the prior could be used to iteratively sample policies for which the likelihood is then being evaluated. This sampling concludes once the agent is sufficiently sure to have sampled enough policies to estimate the posterior accurately.

Mathematically, this process can be described using Markov chain Monte Carlo (MCMC) methods, specifically a modified independent Metropolis-Hastings algorithm which yields a Bayesian independence sampler. Here, instead of using a conditional proposal distribution, a sample policy

π^* is drawn from the prior over policies $p(\pi)$

$$\pi^* \sim p(\pi) \tag{4}$$

$$\rho = \min\left\{\frac{p\left(\pi^*|R\right)p\left(\pi_{n-1}\right)}{p\left(\pi_{n-1}|R\right)p\left(\pi^*\right)}, 1\right\} = \min\left\{\frac{p\left(R|\pi^*\right)}{p\left(R|\pi_{n-1}\right)}, 1\right\} \tag{5}$$

$$\pi_n = \begin{cases} \pi^* & \text{with prob. } \rho \\ \pi_{n-1} & \text{with prob. } 1-\rho \end{cases} \tag{6}$$

after which the likelihood $p\left(R|\pi^*\right)$ of the sample π^* is evaluated, and the sample is accepted or rejected into the chain based on the ratio ρ of the likelihoods of the current and the previous sample (***Equation 5***,***Equation 6***). The samples in the chain constitute i.i.d. drawn samples from the posterior over policies. Note that in this algorithm, sampling noise is determined by the prior over policies, from which new policy samples are drawn. If the prior is very pronounced and many policies have values close to zero, mostly the same few samples will be drawn, making the sampling process less noisy. Conversely, for a flat prior, many different policies are drawn and the sampling process is rather noisy.

Since the posterior being estimated is a categorical distribution, and the samples are i.i.d. drawn, we can infer the parameters ϑ of the sampled distribution $p\left(\pi|\vartheta\right)$ from the entries of the chain, using a Dirichlet prior $p\left(\vartheta|\eta\right)$. To model reaction times, we propose that the sampling concludes once a sufficient level of certainty about the distribution and therefore the best action sequence is reached. To achieve this we use the Dirichlet distribution $p\left(\vartheta|\eta\right)$, whose entropy encodes how certain one can be of having found the true parameters of the distribution to be estimated, where a lower entropy corresponds to more certainty. Hence we use a threshold value $H_{thresh} = H_{init} + \left(H_{init} + 1\right) * f$ of the Dirichlet entropy $H\left[p\left(\theta|\eta\right)\right]$ as a stopping criterion for the sampling, where the free parameter $f \in (0, \infty)$ relates to the initial entropy H_{init} (see Section Value-based decision making in a grid world for influences of f on reaction times). We define the threshold value depending on the initial entropy because for continuous distributions, such as the Dirichlet distribution, entropy may become negative and the initial entropy takes that into account.

Once the entropy has fallen below this threshold, the last sample in the chain determines which policy is executed, and the number of samples the chain $N_{samples}$ required before finishing is taken as an analogon of the reaction time. The resulting algorithm is shown in Algorithm 1 The parameter f determines how much lower than the initial entropy the current entropy has to become, before sampling concludes, and additional implements a constant value, in case the initial entropy is 0. This way, the parameter f will allow us to up- and down-regulate sampling duration and reaction times. Additionally, the shapes of the input distributions influence mean reaction times. As a rule of thumb, under this algorithm sampling concludes earlier and reaction times are faster, the more pronounced prior, likelihood, and the corresponding posterior over policies are. Vice versa, for very uncertain distributions, e.g. when the goal is unknown or uncertain, sampling continues

longer and reaction times become slower.

Data: prior $p(\pi)$, likelihood $p\left(R|\pi\right)$
Result: sample $\pi_{execute}$, number of samples N_{sample}
initialize Dirichlet counts $\eta_i = 1$
initialize entropy $H_{curr} \leftarrow H\left[p\left(\vartheta|\eta\right)\right]$
$n \leftarrow 0$
draw $\pi_0 \sim p(\pi)$
while $H_{curr} > H_{thresh}$ **do**
  $n \leftarrow n + 1$
  draw $\pi^* \sim p(\pi)$
  calculate acceptance probability $\rho = \min\left\{\frac{p(R|\pi^*)}{p(R|\pi_{n-1})}, 1\right\}$
  draw random number $r \in [0, 1]$
  if $r < \rho$ **then**
    $\pi_n \leftarrow \pi^*$
  else
    $\pi_n \leftarrow \pi_{n-1}$
  end
  $\eta_{\pi_n} \leftarrow \eta_{\pi_n} + 1$
  $H_{curr} \leftarrow H\left[p\left(\vartheta|\eta\right)\right]$
end
$N_{samples} \leftarrow n$
$\pi_{execute} \leftarrow \pi_n$

Algorithm 1: Action selection and reaction time algorithm

To describe measured reaction time data, we assume that the true reaction time in milliseconds is linearly related to the number of samples by

$$RT = t_{nd} + t_{sample} N_{samples} \tag{7}$$

multiplying the number of samples $N_{samples}$ with a sampling time t_{sample}, and adding a non-decision time t_{nd} which is due to perceptual processes and loading of information. In the simulations of the flanker and task switching tasks below, we set $t_{sample} = 0.2ms$ and $t_{nd} = 100ms$, to map the number of samples to reaction times below $1000ms$.

Results

We will first illustrate properties of the agent by showing reaction time effects and choice behavior during a sequential value-based decision toy task, where goal-directed action sequences are learned and executed. We specifically show how an agent starts learning rewards and becomes faster in its actions. To focus on the reaction times effects, we keep these simulations deliberately simple and let an agent learn paths to goals in a so-called grid world environment, as used before in theoretical neuroscience, e.g. (***Doya et al., 2002***; ***Schwöbel et al., 2018***; ***Blakeman and Mareschal, 2020***). We go on to show that the information processing principle underlying the prior-based contextual control model can also be applied to qualitatively explain reaction time changes in standard experimental cognitive control tasks. To do this we adapt three key parameters of the model: the initial setting of the prior hyper-paramaters, the expectation of the likelihood of a context change, and the cue perception uncertainty (see Section). We use two different tasks, the Eriksen flanker task which is typically interpreted to measure the inhibition aspect of cognitive control, and Task Switching, which is usually interpreted to measure the cost of switching.

Value-based decision making in a grid world

In this section, we want to present key properties of the prior-based contextual control agent, as well as the novel reaction time algorithm, by showing we can replicate predicted experimental effects, as well as make novel predictions about how prior learning biases action selection and affects

reaction times. Importantly, the agent evaluates behavior based on policies (action sequences) which allows us to not only show behavior and reaction times in single trial experiments, but when an agent learns sequential behavior in a multi-trial simulated experiment. To show these points and present the model in a didactic fashion, we use as a first step a simple and rather artificial value-based decision making toy experiment.

For simulations, we set up a simple grid world (***Figure 1***A). The grid consists of four rows by five columns yielding 20 grid cells. Agents start in the lower middle cell in position 3 (brown square) and have a simple task: to navigate to either one of the two goal positions at cells 15 (blue square) and 11 (green square) while learning their grid world environment. Although the task would not be difficult for a human participant, this task gives us plenty of opportunity to illustrate how the model operates as a value-based decision maker and thereby generates choices and reaction times. In each cell, the agent has three options: move left, up and right. The tasks consists of 200 so-called miniblocks, where one miniblock consists of four trials. In each miniblock, the agent will start in cell 3 and is given the task to move to the indicated goal (either 1 or 2) within the four trials. During the first 100 miniblocks, goal 1 is active, and goal 2 is active in miniblocks 101 - 200. These two phases constitute two distinct contexts as they have different action–outcome contingencies. The agent is not informed about what cells give reward but has to find out by trial and error, inferring the contingencies of the current context.

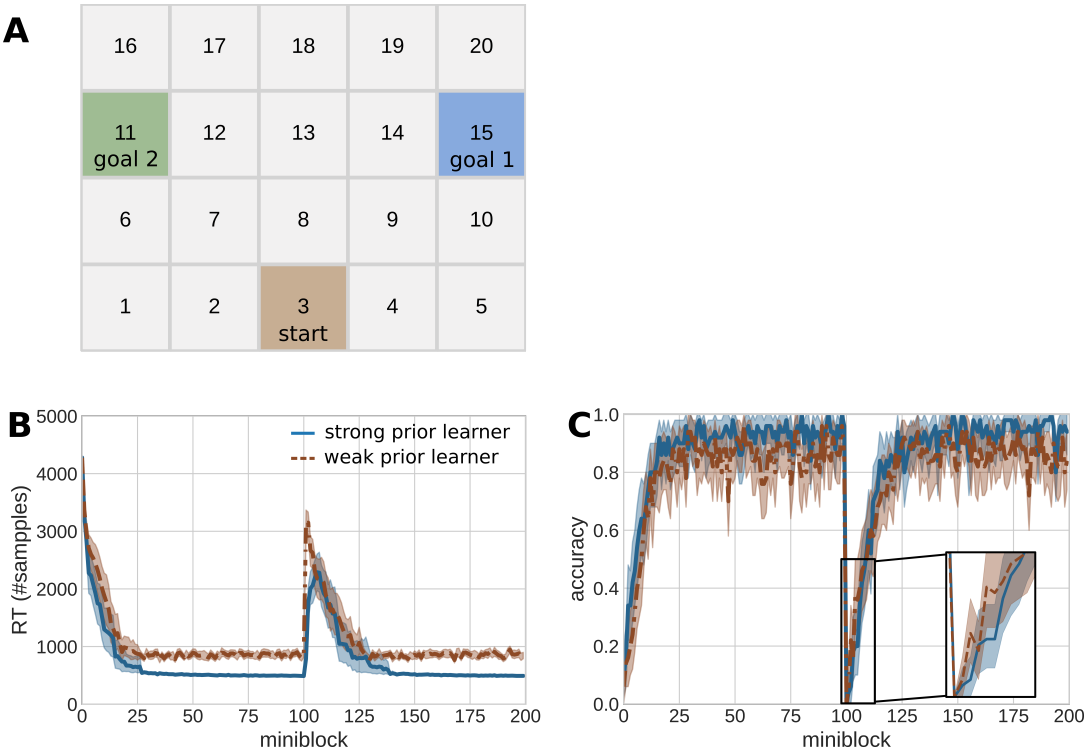

Figure 1. Reaction times and behavior in a sequential instrumental learning task A: The grid world is an environment with 20 states. In each miniblock of four trials, the agent starts in the brown square (cell 3) and has to navigate to either goal 1 (blue square, cell 15) or goal 2 (green square, cell 11). The agent can use either of three actions in each step in the miniblock: left, up, and right. The experiment consist of 200 miniblocks. In the first 100 miniblocks, goal 1 is active, and in the second 100 miniblocks, goal 2 is active. **B**: Reaction times (as number of samples) of the first action in each miniblock over the course of the simulated experiment. The solid blue line shows the mean reaction time of 50 strong prior learning agents ($h = 1.0$). The dashed brown line shows the mean reaction time of 50 weak prior learning agents ($h = 0.001$). The shaded areas indicate a confidence interval of 95%. **C**: Accuracy in the same experiment, colors as in B. The accuracy was calculated as the percentage of miniblocks where agents successfully navigated to the currently active goal (goal 1 in trials 1-100, goal 2 in trials 101-200).

Reaction times and learning

Here we show that reaction times in the beginning of the experiment, as well after a context change, are high and decrease as an agent familiarizes itself with the environment. Additionally, we show that, as expected, how strongly an agent learns a prior over policies, has an effect on how fast reaction times decrease during the transition from goal-directed to more automatized behavior.

To do this, we divided up the agents to be simulated into 50 weak prior learner agents ($h = 0.001$) (see Section Free parameters and simulation setups, and Supplementary files S1) and 50 strong prior learning agents ($h = 1.0$). Note that the weak prior learners adjust their prior over policies so slowly that the prior over policies plays effectively almost no role for action selection. In our simulations, we use the reaction time of the first action as an indicator how fast an agent decided to use a specific sequence of four actions to reach its goal. ***Figure 1***B shows the mean reaction times per miniblock, for strong and weak prior learners. As expected, both agent types, in the beginning of learning how to reach the goal, have slow reaction times. This is because the agent does not know where the goal is yet, which means that all policies have equal value and hence sampling takes longer. These reaction times decrease massively within the first 25 miniblocks, as agents become more certain about the goal location. Strong prior learners additionally learn a pronounced prior over policies (see Section Free parameters and simulation setups, and Supplementary files S1), thereby confining their action selection strongly, and displaying faster reaction times than the weak prior learners. The reaction times of both agent types converge after around 50 trials to stable values, where strong prior learners have generally larger and more variable reaction times.

After trial 100, the contingencies of the environment change and goal 2 becomes active. This means that agents have to infer a new context and learn new goal-directed action sequences. This switching and learning of a new context is clearly expressed in a large increase in reaction times for both agent types, see ***Figure 1***B. As in the first context, both agent types learn the new goal location within the first 25 trials and the mean reaction times decrease again to a stable value.

Figure 1C shows the accuracy of the two agent types. At first, the accuracy increases for both the weak and strong prior learner, as they learn where the goal location is. Once this knowledge and reaction times have stabilized, the strong prior learner navigates to the goal more reliably, leading to an increased accuracy compared to the weak prior learner. This indicates that a strong prior in a stable context not only helps to make faster decisions, but also to choose policies that lead to the goal with more reliability. The accuracy drops to zero as goal 1 is deactivated and goal 2 pays out a reward. The weak prior learner has little a priori tendency to repeat any policies, so it is earlier able to reemploy trial and error and find the new goal location, achieving an average accuracy of 0.164 in trials 101-105. The strong prior learner on the other hand is influenced by the previous prior for 5 trials after the goal switch, which leads to it not becoming as slow with action selection, but also finding and navigating to the new goal later, achieving only an accuracy of 0.096 in trials 101-105.

Time pressure

How does one model time pressure of a decision in the proposed framework? To show this, we analyze the influence of the stopping criterion factor f, see Section Reaction time algorithm. This factor determines how long an agent will sample before committing to a policy, and it therewith up- and down-regulates reaction time means. Even though the stopping criterion factor f is an internal variable, it can be up- or down-regulated in response to external time pressure, allowing us to model time pressure when making a decision. Setting the parameter to a high value means that an agent can sample for a long time, until it is sufficiently certain it knows how successful different policies are expected to be. Setting the factor to a low value, indicates time pressure and a short deadline for making a decision. Values of this parameter not only influence reaction times, but also what policies an agent will choose.

We illustrate this in ***Figure 2***A by showing the Kullback-Leibler (KL) divergence of sampled policies in the Markov chain to the prior and the posterior as a function of the factor f. One can see for

a low factor, i.e. a short deadline, that an agent will rather choose a policy from the prior (low KL divergence to the prior), which corresponds to a repetition-based heuristic of behavior or a habitual choice. For long factors and long deadlines, the agent will choose accurately from the posterior (low KL divergence to the posterior), which balances the a priori heuristic with the goal-directed value in the likelihood.

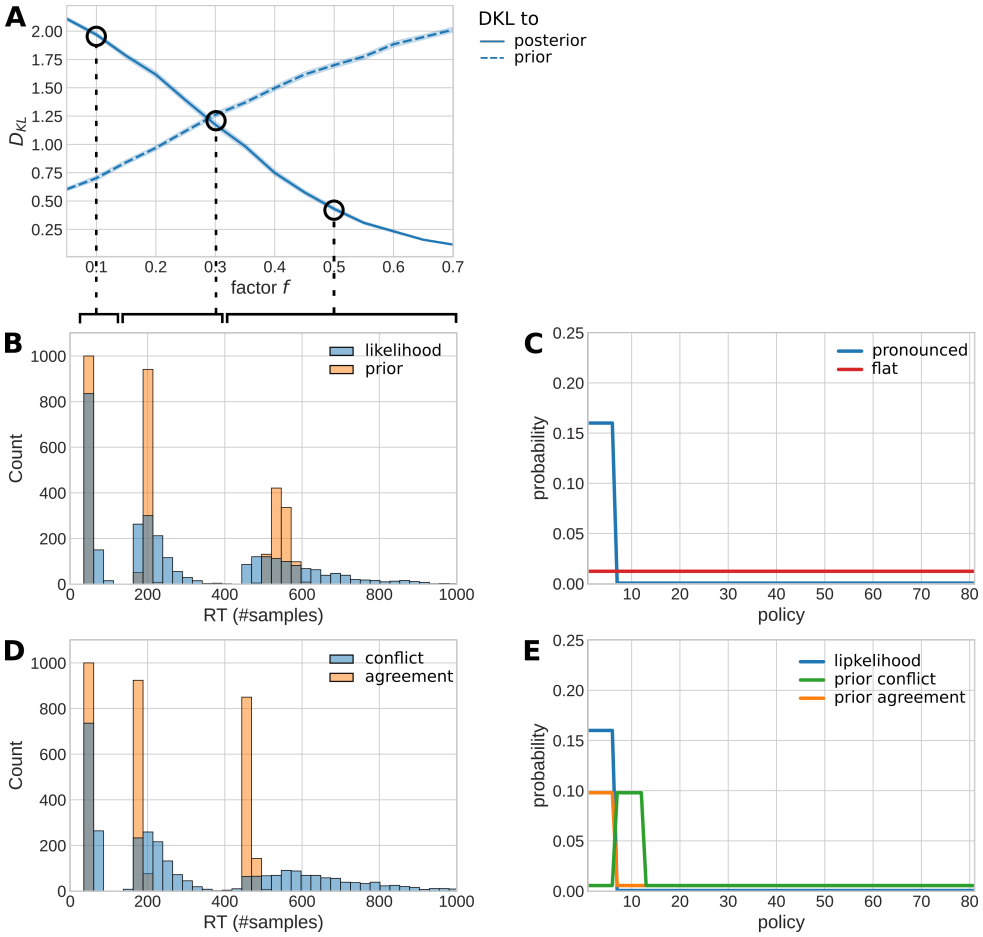

Figure 2. Properties of the RT algorithm This figure illustrates properties of the reaction time algorithm for exemplary values of prior and likelihood that were chosen similar to values which arise in the grid world. **A**: Kullback-Leibler divergence between the samples in the Markov chain and the prior over policies (dashed line) and the true posterior over policies (solid line), as a function of the factor f in the stopping criterion $H_{thresh} = H_{init} + (H_{init} + 1)f$. **B**: RT histograms for different factors f: 0.1 (left), 0.3 (middle), 0.5 (right). The distributions from C were used as input. The orange histograms are from a setting where only the prior was pronounced, but the likelihood was flat, which corresponds to purely prior-based behavior. The blue histograms were from a setting where only the likelihood was pronounced, but the prior was flat, which can be interpreted as purely reward-based behavior. The histograms were produced from 1,000 samples for each f and each combination of prior and likelihood. **C**: Distributions used as input for B. The blue line shows an exemplary pronounced distribution over the 81 policies in the grid world (***Figure 1***A), which is for example found in likelihood after an agent learned the goal location. $81 = 3^4$ is the number of possible combinations of 3 actions that can be used in 4 time steps. Due to the goal position, there are 6 policies that can lead to each goal. Here, they are each assigned a value of ca 0.16, while all other policies have a value close to 0. The red line shows a flat distribution. **D**: RT histograms with the distributions from E as the input, as in B for three different values of f. The orange histograms correspond to RTs when prior and likelihood were in agreement, and the blue histograms correspond to prior and likelihood being in conflict. **E**: Distributions that were used as input for D. The orange and green lines show the distribution of the prior when it was in agreement and conflict, respectively. The numbers were also taken from intermediate values during prior learning in the grid world. The blue line shows the values of the likelihood in this setting.

Goal-based and prior-based behavior

Note that priors over policies play the role of a memory of policies that has been used before in the same context. Here we ask the question how these priors influence reaction times. In what follows, we will show that the modelling framework recreates the known log-normal distribution of

reaction times, as well how the factor f and different combinations of prior and likelihood change the resulting distributions.

Figure 2B shows reaction time distributions for three different exemplary values of the factor f. All distributions exhibit the classical log-normal shape[2], where the distributions for higher factors have a larger mean and a larger variance and tail. An interesting question is how reaction times of an agent which solely relies on its learned prior over policies (orange distributions), differ from reaction times of an agent which, conversely, relies only on the reward-based likelihood (blue distributions). Prior and likelihood shapes are shown in ***Figure 2***C. In both cases, the distributions have a similar mean, but the shapes of the distributions differ substantially. The purely prior-based distributions have a lower variance and no tail, while the purely reward-based distributions have a large variance and are clearly right skewed.

Conflicts

A well-known phenomenon in reaction time experiments is that reaction times are longer for situations where prior information is in conflict with incoming sensory input. We can model this in a straightforward fashion simply by providing the agent with a prior that prefers action sequences which are not supported by the likelihood, see ***Figure 2***E. For example, this conflicting situation arises when the goal switched at trial 101, see ***Figure 1***B, when the agent has learned a prior which is no longer useful for the changed goal location.

We show the resulting reaction time distributions for agreeing and conflicting settings in ***Figure 2***D. Indeed, agreeing distributions lead to reaction time distributions that have a lower mean and variance in contrast to conflicting distributions. This shows that we would expect larger reaction times in conflict situations as agents resolve the conflict, while we see decreased reaction times when both prior and likelihood are in agreement.

Flanker task

The Eriksen flanker task is a widely used behavioral task to measure inhibition under response conflicts (***Gratton et al., 2018***). Typically, in this task participants learn a stimulus-response mapping where one or two stimuli correspond to pressing one key, e.g. right, while one or two different stimuli correspond to pressing another key, e.g. left. The stimulus determining which answer is correct is typically shown in the middle of the screen. Conflict is introduced by showing distractor stimuli (flankers) surrounding the relevant stimulus in the middle. The distractors are chosen to be one of the stimuli that also indicate correct and incorrect responses. This induces congruent trials where the distractors indicate the same key as the relevant stimulus, and incongruent trials where the distractors indicate the other key.

It is typically found that RTs are increased while accuracies are decreased in incongruent trials compared to congruent trials. The classical explanation of this effect is that, early in visual perception, all stimuli are processed in parallel, while perception focuses on the relevant stimulus in a later phase (***Gratton et al., 1992***). Here, we want to show that the flanker effects can be understood in terms of the proposed prior-based contextual control model, where we do not model the perceptual process explicitly, but interpret flankers in terms of context cues and task stimuli in terms of goal-directed information. Concretely, we naturally interpret the experiment such that the flanker distractor stimuli are perceived as context cues which indicate what the correct response would be. To realize this in the model, we use an a priori inference scheme by setting up the hyperparameters of the prior over actions (see Section Free parameters and simulation setups, and Supplementary files S1) to encode pre-existing cue–response associations, which map a response to a specific context, effectively implementing a priori flanker–response associations. During the experiment, this results in a quick loading of the respective prior over this action, which is supposed to facilitate fast action selection. In an incongruent trial, this prior is in conflict with the actual goal-directed

[2]Log-normality was tested using the python scipy normaltest function on the log of the RT in number of samples. All distributions pass the normal test with $p < 0.01$

stimulus encoded in the likelihood. Additionally, this inference machine is by design subject to a small forgetting factor (see and Supplementary files S1) which leads to an updating of the flanker–response associations during the course of the experiment. In our simulations, we used a flanker task version with four stimuli.

Conditional accuracy function

One typical finding in flanker tasks is the conditional accuracy function (***Figure 3***A) (***Stins et al., 2007***). In congruent trials, responses preceded by long as well as short RTs are correct to a high degree. In incongruent trials, responses preceded by short RTs are most likely incorrect with the accuracy dropping below chance. The more the RT increases, the more the accuracy of responses increases. ***Figure 3***B shows the averaged accuracy function of 50 simulated agents which shows the typical dip below chance for low RTs in incongruent trials. In the model, this effect comes from the fact that for lower RTs, choices are more often made in accordance with the prior, while for long RTs, choices are mostly made in accordance with the posterior (see also ***Figure 2***A).

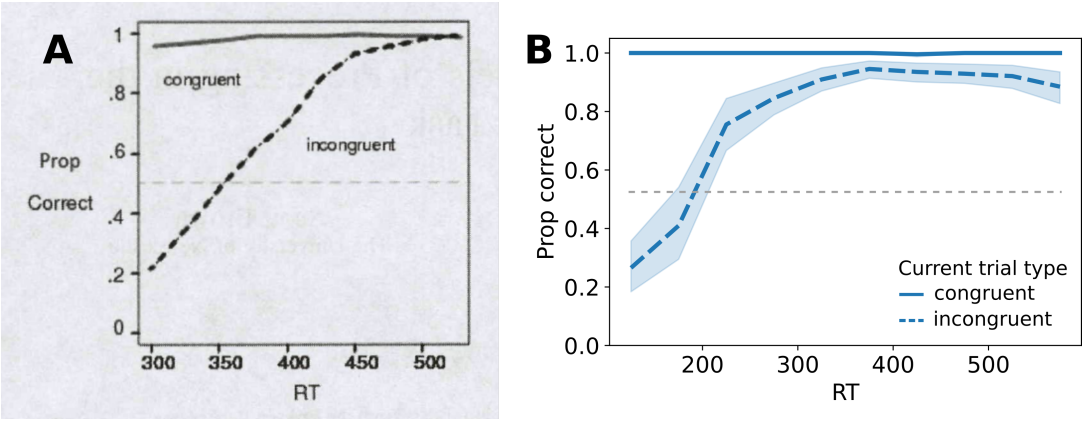

Figure 3. Conditional accuracy function A: Idealized accuracy function in the flanker task, figure taken from (***White et al., 2012***). The solid and dashed lines show the proportion of correct responses for congruent and incongruent trials, respectively, as a function of reaction time. The grey dashed line indicates chance level. **B**: Simulated accuracy function, the line styles are as in A. The lines indicate the mean proportion of correct responses of 50 agents. The shaded area shows the confidence interval of 95%. The proportions were calculated by binning reaction times.

Gratton effect

Another typical finding in the flanker task is the so-called Gratton effect (***Figure 4***A) (***Gratton et al., 1992***). Here, mean reaction times decrease in the second consecutive trial of the same type (congruent vs incongruent). Usually, the Gratton effect is interpreted as being due to different degrees of recruitment of cognitive control depending on the previous trial type. According to our model, the Gratton effect would rather be a sequential effect, which is due to strengthened or weakened associations between the distractor (context) and the response. A congruent trial would strengthen the association, making prior and likelihood more or less similar in a following congruent or incongruent trial. This in turn de- or increases reaction times. In ***Figure 4***B, we show that we can replicate the Gratton effect using this mechanism.

In order to demonstrate that the setup of the prior, the experience-dependent updating of the prior, and importantly, its interplay with the context are the key machinery which allows the prior-based contextual control model to simulate typical flanker effects, we show in the Supplementary files S2 that leaving out any of these three components will drastically change the conditional accuracy function and nullify the Gratton effect.

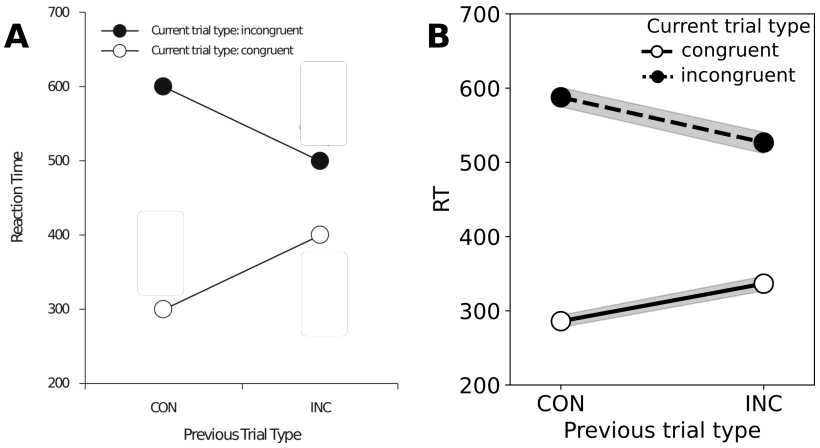

Figure 4. Gratton effect A: Typical Gratton effect findings in the flanker task, figure taken from (***Davelaar and Stevens, 2009***). The x-axis shows the previous trial type, either congruent (CON) or incongruent (INC). The circles indicate the current trial as either incongruent (black) or congruent (white). Reaction times on the y-axis differ between congruent and incongruent trials, and depending on the previous trial type. **B**: Simulated Gratton effect. The axes are as in A, and the dashed line indicates a current incongruent trial, and the solid line a current congruent trial. The shaded areas show a confidence interval of 95%.

Task switching

In a typical task switching task (***Kiesel et al., 2010***; ***Monsell, 2003***), participants are presented with two different response rules, for example participants must indicate whether a number is even or odd, and whether a letter is a vowel or a consonant. In each trial, the current task set is cued, and a stimulus is presented with features from both task sets, such as a letter and a number. The participant has to respond to the stimulus under the current task set. Due to the stimulus consisting of both features, trials can be congruent or incongruent: In congruent trials both features require the same response, and in incongruent trials the two features require different responses. In this task, subjects experience two sources of uncertainty: The context cue may be perceived or processed in a noisy manner, which is also influenced by the cue presentation time, i.e. how long the task cue is visible before a response is warranted; and uncertainty about the upcoming task and context, and how often this changes.

In the proposed model, task switching corresponds to switching between two different contexts with different outcome rules. The agent learns the outcome rules for each context at the beginning of the experiment and later in the experiment loads the task-specific learned rules in response to the task set (context) cue. The agent is set up such that context cues are observed with low but non-zero uncertainty, and with the agents' expectation of context transition probability set to be relatively low as well (see also Section Free parameters and simulation setups). This leads to uncertainty in the context inference and results in the previous context still being loaded to some degree in a switch trial, which decreases as the number of consecutive trials in the same context decreases. The previous context may then introduce conflicts which increase reaction times. Note that here, we use a weak prior learner ($h = 0.001$) to focus less on learning a prior, probably as a human participant would do when switching tasks are expected.

In this section, we show three common findings from the task switching literature (***Kiesel et al., 2010***; ***Monsell, 2003***; ***Steyvers et al., 2019***; ***Jamadar et al., 2015***): (1) decreased reaction times in repetition trials of the same task set, (2) decreased reaction times with longer cue–target intervals, and (3) decreased reaction times with longer response–stimulus intervals.

Repetition trials

In the first trial after a task switch, reaction times typically increase. This finding is typically interpreted as being caused by underlying costs associated with switching the task set. In particular, the

previous task set may interfere with the response to the new task set. Additionally, as in the Flanker task, participants are typically slower in incongruent trials than in congruent trials. Lastly, reaction times typically decrease as more trials in the task have been trained. These results are shown in FIGnuminrunA. FIGnuminrunB shows simulated average reaction times of 50 agents over the course of a 70 trial task switching experiment. As in the experimental findings on the left, shorter periods of training lead to larger reaction times. Both in the experiment and the model, incongruent trials lead to increased reaction times, especially in switch trials. In addition, reaction times decrease as a function of repeated trials.

In the agent, these effects are due to the previous task, i.e. context, lingering because of the remaining uncertainty in the context inference. The longer the same task was active, the lower the remaining uncertainty over tasks, and the less the action-outcome contingencies of the previous task influence the decision. This effect is stronger in conflict trials, as the contingencies of the other task may contradict the contingencies of the current task, which induces uncertainty in the action selection and therewith increases reaction times.

The converse has been found experimentally for response accuracy: Accuracy drops in switch trials, and is lower in incongruent trials (FIGnuminrunC). In FIGnuminrunD, we show the simulated average accuracy. The accuracy difference between congruent and incongruent trials is not quite as large in the simulated data compared to the experimental data, but it is present. As described above, these effects are due to the contingencies of the previous task influencing action selection in proportion to the context uncertainty.

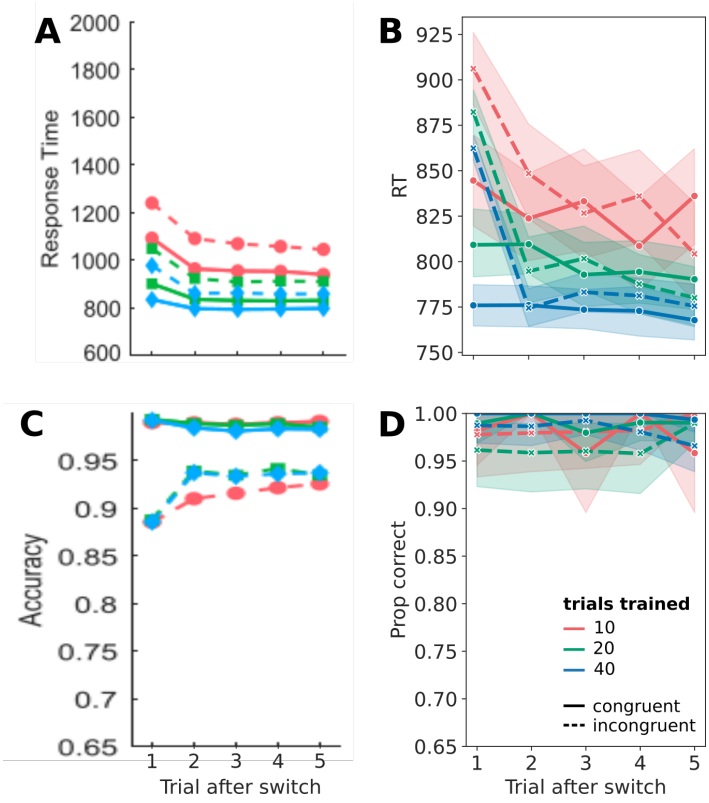

Figure 5. RTs and accuracy as a function of repeating trial after switch **A**: Reaction time in a typical task switching experiment as a function of trial number after the switch, taken from (***Steyvers et al., 2019***). Red lines indicate little previous training (~10 trials), green lines indicate medium training (~20 trials), and blue lines long training (~40 trials). Dashed lines are incongruent trials, and solid lines congruent trials. The shaded areas correspond to a 95% confidence interval. **B**: Simulated average reaction times from 50 agents. Colors as an A. **C**: Response accuracy as a function of trial number after switch from the same experiment as A. Colors as in A. **D**: Simulated average response accuracy from the same simulated experiment as in C. Colors as in A.

To demonstrate that context inference and context–dependent learning are essential for modeling a task switching task, we show in the Supplementary files that removing the context feature leads to vastly different reaction time and accuracy effects compared to those shown here, see Supplementary files S3, which do not resemble the typical task effects anymore.

Cue–target interval

Another well known effect in task switching is the influence of the cue–target interval (CTI) on reaction times (***Jamadar et al., 2015***). The cue–target interval is the time between the presentation of the current task cue and the onset of the stimulus (target) upon which a response has to be made. Longer CTIs allow participants to better process the cue and load the task set before the onset of the stimulus. FIGCTIA shows how reaction times increase with shorter CTI, compared to a single task experiment (***Jamadar et al., 2015***).

To model the CTI, we map shorter CTIs to higher uncertainty when perceiving the context cue that indicates the current task (see Section Free parameters and simulation setups, and Supplementary files S1). FIGCTIB shows average reaction times in switch trials, repeat trials, and trials in a single task experiment as a function of context cue uncertainty. Using this setup, we were able to qualitatively recreate the typical shape of CTI effects. The higher uncertainty during perception and processing of the task cue leads to increased reaction times because the previous task's contingencies have higher influence on action selection.

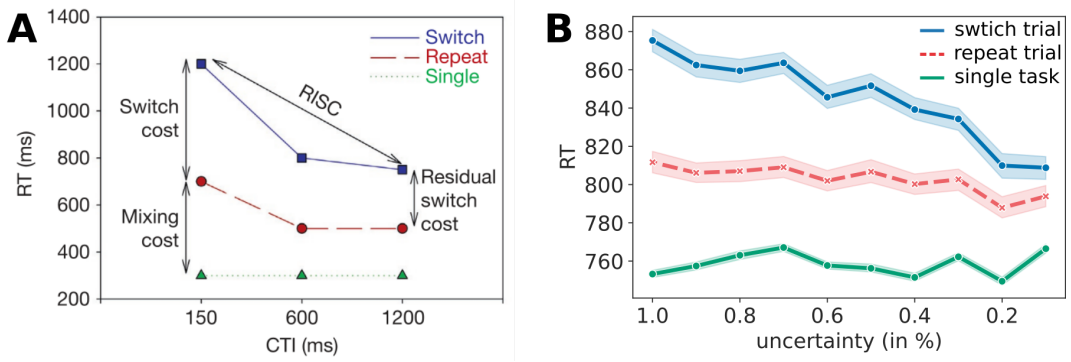

Figure 6. Cue–target interval A: Reaction times in a single task experiment (green), in repeat trials in a task switching experiment (red), and in switch trials (blue), as a function of CTI; adapted from (***Jamadar et al., 2015***). **B**: Mean simulated reaction times from 50 agents, as a function of cue uncertainty, colors as in A. The shaded ares indicate the 95% confidence interval.

Response–stimulus interval

A similar effect has been observed when varying the response–stimulus interval (RSI) (***Monsell, 2003***), which is the time between a response and the next trial. Longer RSIs allow participants to better prepare for a possible switch and as a result decrease reaction times. FIGRTIA shows reaction times in switch and repeat trials as a function of the RSI (***Monsell, 2003***).

We model RSI differences as different levels of an agent's assumption about context change probability in between trials (see Section Free parameters and simulation setups, and Supplementary files S1). The higher the change probability, the easier it is for an agent to infer that the context and task changed, and to load the new task set and respond accordingly. FIGRTIB shows average reaction times as a function of change probability. The same decrease of reaction times with higher change probability emerges, as with the RSI.

Discussion

We have proposed a joint behavioral modeling approach for choices and reaction times in cognitive neuroscience experiments. To model (value-based) choices we use a Bayesian model of

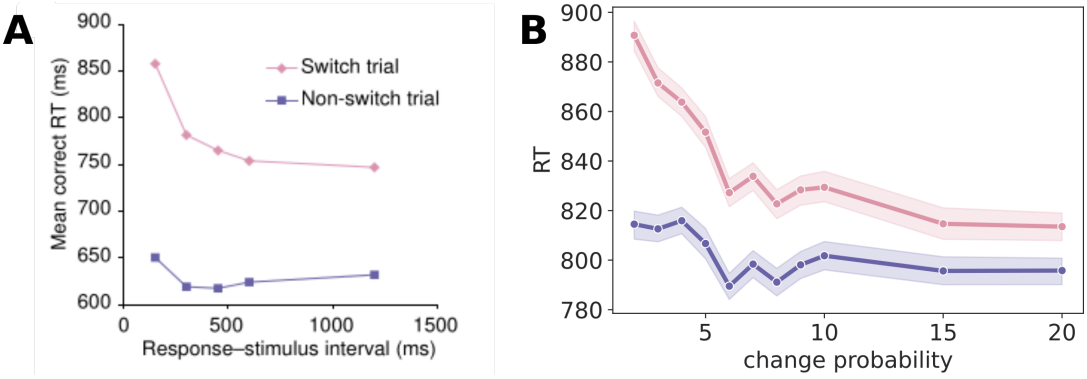

Figure 7. Response–stimulus interval A: reaction times as a function of RSI for switch trials (pink) and repeat trials (purple), taken from (***Monsell, 2003***). **B**: Mean simulated reaction times from 50 agents, as a function of change probability, colors as in A. The shaded areas indicate the 95% confidence interval.

goal-directed behavior with two key components: (i) A prior over actions or policies (sequences of actions) which is learned over time to match the history of past actions (repetition-based prior); and (ii) a hierarchical factorization of latent variables into states and contexts, where contexts determine the state transition and outcome rules of the environment. Additionally, the prior is learned in a context-dependent manner and encodes which policies should be preferred by the agent in the current context. This model is coupled with a reaction time algorithm which is based on a Bayesian independent Markov Chain Monte Carlo (MCMC) sampler which evaluates and selects policies. The MCMC sampler uses prior and likelihood of policies to generate reaction times and to sample policies from the posterior, which are executed.

Using this approach, we showed that the components of this joint model can qualitatively explain several experimental findings in value-based decision tasks as well as typical cognitive control experiments: (1) Decreases in reaction times in a sequential value-based decision task. (2) Classical log-normal shapes of reaction time distributions and prior-based, automatic choices for shorter reaction times. (3) For a flanker task, we replicated the reaction time dependent conditional accuracy function, as well as the sequential Gratton effect. (4) In a typical task switching task, we replicated repetition effects, cue–target interval (CTI) effects, and inter-trial (response–stimulus interval, RSI) effects.

Both the adaptive prior over policies and the hierarchical contextualisation of the generative model are critical for capturing this rather wide range of effects (see also Supplementary files). The context-dependent prior over actions or policies encodes a tendency to repeat specific policies in a specific context. This allowed us to model how behavioral sequences can become more independent of goal representations and thereby automatized. This increases accuracy and decreases reaction times and helps an agent to make faster and more reliable choices. Additionally, this prior encodes cue–response associations which enabled us to model response conflicts in the flanker task, where distractors indicate a different response than the relevant stimulus.

The hierarchical description of the dynamic tasks as context dependent action–outcome contingencies and state transitions furthermore enabled us to model changes in the task structure and dynamics as for example in task switching. As the context itself is a latent variable, the agent forms beliefs over possible contexts and assigns precision to those beliefs. Imprecise beliefs over contexts are critical to modeling effects in a task switching task, as both the effects of the cue–target interval as well as the response–stimulus interval depended on context inference uncertainty.

We interpret our findings such that the general process of perceiving and processing the context cue, inferring the currently active context, loading the respective action–outcome contingencies and prior over actions, and then using these to plan ahead and choosing an action is a common mechanism underlying all tasks shown in this work.

Although the typical cognitive control tasks are usually not interpreted to rely on value-based decision making, the two key components in the model seem to capture some of the underlying cognitive processes. Hence, the proposed modeling approach leads to an interesting view on cognitive control, which is typically interpreted as top-down control (***Botvinick et al., 2001***; ***Miyake et al., 2000***; ***Goschke, 2014***; ***Gratton et al., 2018***). In contrast, in the model, bottom-up inference of the current context plays an important role as well. Here, an inferred context determines not only which task rules currently hold but also which policies should be preferred in the current situation. Consequently, contexts are modeled at a higher level in the hierarchy, relative to actions. This enables the agent to probabilistically infer the high-level state context from its sensory inputs. In this sense, cognitive control, from a modeling perspective, is not only about 'control' but also about 'inference' (***Attias, 2003***; ***Botvinick and Toussaint, 2012***). This 'cognitive inference' determines what rules the agent should currently follow, and which a priori information to use (***Marković et al., 2019***). To an outside observer, some resulting behavior may look like 'top-down control' but may be understood as the agent's recognition of the current situation based on context inference and having learnt how to behave in it, based on previous exposures to the same or similar situations (***Lieder et al., 2018***).

Relation to other joint behavioral models

The classical modeling approach for reaction times is the influential diffusion decision model (DDM) (***Ratcliff, 1978***; ***Boucher et al., 2007***; ***Ratcliff et al., 2016***; ***Forstmann et al., 2016***) which is in cognitive science one of the most established textbook models (***Forstmann and Wagenmakers, 2015***, Chapter 3). It has been successfully applied across many experimental domains, most notably to perceptual decision making where the task rests on perceptual evidence accumulation. DDMs fall under the umbrella of evidence accumulation models which describe the process of action selection and resulting reaction times as a biased random walk with a drift and white noise. This approach has been extended to multi-choice tasks in so-called race diffusion models, where instead of having one accumulator as in the DDM, each available choice option is associated with a different accumulator (***Fontanesi et al., 2019***; ***Miletić et al., 2021***).

The evidence accumulation reaction time modeling approach has been recently combined with reinforcement learning models to provide joint instrumental learning and reaction time models (***Milosavljevic et al., 2010***; ***Pedersen et al., 2017***; ***Fontanesi et al., 2019***; ***Miletić et al., 2021***). Here, usually internal variables from the reinforcement learning agent, particularly expected reward (Q-values) are mapped to variables in the evidence accumulator model, such as the drift rate, e.g. (***Miletić et al., 2021***). Learned values or value differences drive the random walk until a boundary is reached and the respective choice is executed. These models were typically validated in one-step instrumental learning experiments and reproduce the classic log-normal shape of reaction time distributions. However, the resulting modeling approach lacks mechanistic appeal as it is unclear what the white noise component of the DDM represents. As it stands, this noise component in DDMs is a useful modeling device to explain reaction time distributions but it is difficult to link to an underlying generative mechanism.

Consequently, a key difference of the DDM-based approach to our proposed model is that we use sampling not only to provide a way to generate reaction times from the choice values of the value-based model, but importantly to describe a potential mechanism how the inherent noise of the sampling contributes to the actual decision process. Policies are sampled from the prior over policies, which provides a heuristic about which parts of the decision tree should be evaluated first, and which can be ignored. For each sampled policy, the goal-directed value is computed. This way, the prior not only encodes context–response associations but also helps to confine the space of what behavior to evaluate in a goal-directed manner. If the prior is close to zero, as is typically the case for many policies, these are effectively a priori, due to the sampling, excluded, and the prior hence helps to save resources and time when selecting actions. Taken together the sampling describes a mechanistic iteration to test, in a manner informed by the prior, hypotheses about

602 future behavior and which action could be the best. Naturally, we model the reaction time as the
603 duration of this sampling process.

604 The sampling continues until the agent is sufficiently certain that the sampling has yielded
605 enough information on what outcomes to expect. The level of certainty is regulated by the stop-
606 ping criterion factor f. This approach automatically integrates the uncertainty about outcomes into
607 action selection, and is related to optimal stopping problems (***Shiryaev, 2007***). For small factors,
608 sampling stops when the sampled estimate crosses the true posterior of a policy for the first time,
609 while larger factors let the sampling converge longer, thereby leading to more accurate values.
610 Furthermore, certainties of the prior, as well as goal-directed values in the likelihood are reflected
611 in different reaction time distribution shapes and means, for the same factor value. These differ-
612 ences in resulting RT distributions for different prior and likelihood combinations can be linked to
613 between participant RT differences measured in behavioral experiments, and therefore enabling
614 identification of generative models employed by the participants (***Daunizeau et al., 2010***).

615 MCMC was chosen as a sampling algorithm because it is a well-established general method
616 to sample from probability distributions. It has even been argued recently that probabilistic com-
617 putations may be implemented by neurons in the brain via types of MCMC sampling (***Pecevski***
618 ***et al., 2011***; ***Sanborn and Chater, 2016***; ***Aitchison and Lengyel, 2016***). In this article, we focused
619 on describing reaction time variability as caused by sampling in the decision and action selection
620 process. However, it is also possible that additional RT variability may be generated by sampling in
621 other domains, such as in the planning process which can be implemented via sampling (***Browne***
622 ***et al., 2012***; ***Vien et al., 2013***; ***Fountas et al., 2020***), and perceptual processes (***Orbán et al., 2016***;
623 ***Echeveste et al., 2020***). For example in the flanker task, past modeling approaches have focused
624 on perceptual uncertainty to explain reaction time effects, e.g. (***White et al., 2012***) while we fo-
625 cused solely on the decision process itself. It is hence possible that explicitly modeling perceptual
626 effects in the flanker task may further improve the match to observed reaction time distributions
627 and effects.

628 Acknowledgments

629 We thank Maarten Jung for his help with a previous version of the MCMC sampling algorithm as
630 part of his masters thesis (https://nbn-resolving.org/urn:nbn:de:bsz:14-qucosa2-740483).

631 Funding acknowledgements

632 Funded by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft), SFB 940/3
633 - Project number 178833530, and TRR 265 - Project number 402170461, and as part of Germany's
634 Excellence Strategy – EXC 2050/1 – Project ID 390696704 – Cluster of Excellence "Centre for Tactile
635 Internet with Human-in-the-Loop" (CeTI) of Technische Universität Dresden.

636 Additional files

637 Supplementary files

638 • S1: Detailed derivations of the mathematical model.
639 • S2: Flanker features, where parts if the agent's machinery is switched off.
640 • S3: Task switching features, where parts if the agent's machinery is switched off.

641 Code availability

642 The code has been made publicly available on github:
643 https://github.com/SSchwoebel/BalancingControl

644 References

645 **Aitchison L**, Lengyel M. The Hamiltonian brain: Efficient probabilistic inference with excitatory-inhibitory neu-
646 ral circuit dynamics. PLoS computational biology. 2016; 12(12):e1005186.

647 **Attias H**. Planning by probabilistic inference. In: *AISTATS*; 2003. .

Blakeman S, Mareschal D. A complementary learning systems approach to temporal difference learning. Neural Networks. 2020; 122:218–230.

Botvinick M, Toussaint M. Planning as inference. Trends in cognitive sciences. 2012; 16(10):485–488.

Botvinick MM, Braver TS, Barch DM, Carter CS, Cohen JD. Conflict monitoring and cognitive control. Psychological review. 2001; 108(3):624.

Boucher L, Palmeri TJ, Logan GD, Schall JD. Inhibitory control in mind and brain: an interactive race model of countermanding saccades. Psychological review. 2007; 114(2):376.

Browne CB, Powley E, Whitehouse D, Lucas SM, Cowling PI, Rohlfshagen P, Tavener S, Perez D, Samothrakis S, Colton S. A survey of monte carlo tree search methods. IEEE Transactions on Computational Intelligence and AI in games. 2012; 4(1):1–43.

Daunizeau J, Den Ouden HE, Pessiglione M, Kiebel SJ, Stephan KE, Friston KJ. Observing the observer (I): meta-bayesian models of learning and decision-making. PloS one. 2010; 5(12):e15554.

Davelaar EJ, Stevens J. Sequential dependencies in the Eriksen flanker task: A direct comparison of two competing accounts. Psychonomic bulletin & review. 2009; 16(1):121–126.

Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. Model-based influences on humans' choices and striatal prediction errors. Neuron. 2011; 69(6):1204–1215.

Doya K, Samejima K, Katagiri Ki, Kawato M. Multiple model-based reinforcement learning. Neural computation. 2002; 14(6):1347–1369.

Echeveste R, Aitchison L, Hennequin G, Lengyel M. Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference. Nature neuroscience. 2020; 23(9):1138–1149.

Fontanesi L, Gluth S, Spektor MS, Rieskamp J. A reinforcement learning diffusion decision model for value-based decisions. Psychonomic bulletin & review. 2019; 26(4):1099–1121.

Forstmann BU, Ratcliff R, Wagenmakers EJ. Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. Annual review of psychology. 2016; 67:641–666.

Forstmann BU, Wagenmakers EJ. Chapter 3. In: An introduction to model-based cognitive neuroscience Springer; 2015. .

Fountas Z, Sajid N, Mediano PA, Friston K. Deep active inference agents using Monte-Carlo methods. arXiv preprint arXiv:200604176. 2020; .

Friston K. The free-energy principle: a rough guide to the brain? Trends in cognitive sciences. 2009; 13(7):293–301.

Friston K. The free-energy principle: a unified brain theory? Nature reviews neuroscience. 2010; 11(2):127–138.

Friston K, FitzGerald T, Rigoli F, Schwartenbeck P, Pezzulo G, et al. Active inference and learning. Neuroscience & Biobehavioral Reviews. 2016; 68:862–879.

Friston K, Rigoli F, Ognibene D, Mathys C, Fitzgerald T, Pezzulo G. Active inference and epistemic value. Cognitive neuroscience. 2015; 6(4):187–214.

Goschke T. 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. 2014; 23(S1):41–57.

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

Gratton G, Cooper P, Fabiani M, Carter CS, Karayanidis F. Dynamics of cognitive control: Theoretical bases, paradigms, and a view for the future. Psychophysiology. 2018; 55(3):e13016.

Jamadar S, Thienel R, Karayanidis F. Task switching processes. Brain mapping: An encyclopedic reference. 2015; 3:327–335.

Kiesel A, Steinhauser M, Wendt M, Falkenstein M, Jost K, Philipp AM, Koch I. Control and interference in task switching—A review. Psychological bulletin. 2010; 136(5):849.

694 **Kolling N**, Wittmann M, Rushworth MF. Multiple neural mechanisms of decision making and their competition
695 under changing risk pressure. Neuron. 2014; 81(5):1190–1202.

696 **Kozak MJ**, Cuthbert BN. The NIMH research domain criteria initiative: background, issues, and pragmatics.
697 Psychophysiology. 2016; 53(3):286–297.

698 **Lieder F**, Shenhav A, Musslick S, Griffiths TL. Rational metareasoning and the plasticity of cognitive control.
699 PLoS computational biology. 2018; 14(4):e1006043.

700 **Marković D**, Reiter AM, Kiebel SJ. Predicting change: Approximate inference under explicit representation of
701 temporal structure in changing environments. PLoS computational biology. 2019; 15(1):e1006707.

702 **Miletić S**, Boag RJ, Trutti AC, Stevenson N, Forstmann BU, Heathcote A. A new model of decision processing in
703 instrumental learning tasks. Elife. 2021; 10:e63055.

704 **Milosavljevic M**, Malmaud J, Huth A, Koch C, Rangel A. The drift diffusion model can account for value-based
705 choice response times under high and low time pressure. Judgment and Decision Making. 2010; 5(6):437–
706 449.

707 **Miyake A**, Friedman NP, Emerson MJ, Witzki AH, Howerter A, Wager TD. The unity and diversity of executive
708 functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. Cognitive psy-
709 chology. 2000; 41(1):49–100.

710 **Monsell S**. Task switching. Trends in cognitive sciences. 2003; 7(3):134–140.

711 **Orbán G**, Berkes P, Fiser J, Lengyel M. Neural variability and sampling-based probabilistic representations in
712 the visual cortex. Neuron. 2016; 92(2):530–543.

713 **Pecevski D**, Buesing L, Maass W. Probabilistic inference in general graphical models through sampling in
714 stochastic networks of spiking neurons. PLoS computational biology. 2011; 7(12):e1002294.

715 **Pedersen ML**, Frank MJ, Biele G. The drift diffusion model as the choice rule in reinforcement learning. Psy-
716 chonomic bulletin & review. 2017; 24(4):1234–1251.

717 **Ratcliff R**. A theory of memory retrieval. Psychological review. 1978; 85(2):59.

718 **Ratcliff R**, Smith PL, Brown SD, McKoon G. Diffusion decision model: Current issues and history. Trends in
719 cognitive sciences. 2016; 20(4):260–281.

720 **Sanborn AN**, Chater N. Bayesian brains without probabilities. Trends in cognitive sciences. 2016; 20(12):883–
721 893.

722 **Schwöbel S**, Kiebel S, Marković D. Active inference, belief propagation, and the bethe approximation. Neural
723 computation. 2018; 30(9):2530–2567.

724 **Schwöbel S**, Marković D, Smolka MN, Kiebel SJ. Balancing control: a Bayesian interpretation of habitual and
725 goal-directed behavior. Journal of Mathematical Psychology. 2021; 100:102472.

726 **Shahar N**, Hauser TU, Moutoussis M, Moran R, Keramati M, Consortium N, Dolan RJ. Improving the reliabil-
727 ity of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-
728 diffusion modeling. PLoS computational biology. 2019; 15(2):e1006803.

729 **Shiryaev AN**. Optimal stopping rules, vol. 8. Springer Science & Business Media; 2007.

730 **Steyvers M**, Hawkins GE, Karayanidis F, Brown SD. A large-scale analysis of task switching practice effects
731 across the lifespan. Proceedings of the National Academy of Sciences. 2019; 116(36):17735–17740.

732 **Stins JF**, Polderman JT, Boomsma DI, de Geus EJ. Conditional accuracy in response interference tasks: Evidence
733 from the Eriksen flanker task and the spatial conflict task. Advances in cognitive psychology. 2007; 3(3):409.

734 **Vien NA**, Ertel W, Dang VH, Chung T. Monte-Carlo tree search for Bayesian reinforcement learning. Applied
735 intelligence. 2013; 39(2):345–353.

736 **White CN**, Brown S, Ratcliff R. A test of Bayesian observer models of processing in the Eriksen flanker task.
737 Journal of Experimental Psychology: Human Perception and Performance. 2012; 38(2):489.