University of Nottingham

Department of Psychology

Master Thesis

submitted for the degree of

MSc Computational Neuroscience, Cognition and AI

Measuring the degree of model-based control

Based on a study conducted by Kool et al., (2018)

by

Matthieu Pâques
Module: PSGY4012 UNUK
Student Id.: 20314310

Supervisor: Carolina Feher da Silva
Submission Date: October 8, 2021

Abstract

Human decision making is thought to rely on two distinct systems known as the model-based (MB), responsible for planning, and the model-free (MF), embodying habits. To retrieve the strategy used by participants, many studies have made an extensive use of a dual system mixing the two models. However, the evidences in favour of a MF are either weak or controversial. Moreover, we show that the mixing weights parameters of the dual model are subject to noisy estimation. We present a new approach relying on the likelihood of a MB agent to assay the degree of MB control. Unlike the dual approach, the new method proves to be reliable and valid to compare between participants in a multistage task. To allow a comparison between the stages of the task, new metrics are designed using the MB likelihood. However, despite encouraging reliability results, none of the aforementioned methods is capable of comparing the degree of MB control across the stages of a task.

Contents

List of Tables

List of Figures

1 Introduction

Being computational or mental, decision-making is subject to a trade-off between accuracy and computational cost. A complex planning of the choice is made at the expense of a longer reaction time or higher mental effort. It is thought that humans would rely on different systems as a function of the available cognitive resources. The allocation of cognitive resources to a task might depend on the expected rewards but also on fatigue, stress or constraints on the response time. One can distinguish two main systems competing for the control of behaviour. The habit system, fast and automatic, is also less accurate. The goal-directed system, slow and deliberative, often reaches higher accuracy (Kool et al. 1985, Kahneman 2003). The cost benefit trade-off consists in the use of the most accurate model when cognitive resources are available and the use of the habit system otherwise. Using a Reinforcement Learning (RL) formalization, the habit system and the goal-directed system would correspond respectively to the model-free (MF) agent and the model-based (MB) agent. The MF agent possesses no knowledge of the environment and simply reproduces the actions that previously gave the maximum reward. The need of previous experience to update its value function makes the MF model inflexible to sudden changes in the environment. On the other hand, the MB model masters the structure of the environment and relies on it to plan its decisions. A change of environment can be taken into account by modifying its internal representation of the task at the expense of computational effort. Thus the MB model is more accurate and flexible but also more computationally expensive. The hybrid system combining the model-free and model-based agents matches the behaviour of Human participants in different tasks (Gillan et al. 2015, Otto et al. 2015). A functional Magnetic Resonance Imaging study also detected a neural signature for the prospective evaluation of a choice consistent with a MB approach (Doll et al. 2015). However, recent studies questioned this well-established dual systems. First, the equivalence between the habit system and a model-free system isn't consensual. Miller et al. have designed an alternative model functioning without outcome encoding and able to account for habit behaviour (Miller et al. 2019). Instead of estimating the expected rewards, habits develop by strengthening the recent actions. This model was able to explain several key habit behaviours, as perseveration despite reward devaluation. Second, unlike the MB system, no neurological evidences were found to support the existence of the MF system. The hippocampus is believed to have a key role in decision making. However, rats with inactivated hippocampus revealed no impairment in their use of a MF strategy whilst their reliance on the planning system was reduced (Miller et al. 2017). Up to date, no other brain region has proved to support an internal representation of the MF system. Third, a method fitting Human's data into two well-defined models omits the contingency of Human decision making. Human decision making is a complex process to study as it follows sometimes surprisingly sub-optimal policies, matching probabilities instead of maximising the reward expectation (Vulkan 2000, Newell & Schulze 2000). Humans have also the tendency to distinguish pattern in noise, making them act unexpectedly even in the simpler tasks (Huettel et al. 2002). Misconceiving a task, by irrelevantly giving values to the object location or the decision keyboards, can also disrupt the identification of the strategy. Feher da Silva et al. shown that a pure model-based agent relying on a misconceived task can mimic the existence of a mixture of model-based and model-free agents (Feher da Silva & Hare 2020). Importantly, identifying the decision strategy has established itself as a promising tool in the diagnosis of behavioural disorders (Gillan et al. 2016). Patients with obsessive-compulsive disorders (OCD) seemed to show weaker goal-directed performance compared to healthy patients (Gruner et al. 2016). Consistent with an extensive reliance on the habit system, OCD subjects suffer from repetitive thoughts and behaviours. A similar bias toward a model-free strategy is noticeable on people presenting food and drug addiction (Voon et al. 2015) (Redish et al. 2008).

The objective of this work is to propose new methods able to account for the degree of model-based control in a multi-stage task. First, we will investigate the ability of the dual system (MBMF) to assess the parameters of a simulated strategy. Second, we will search for alternative methods to the dual system to assay the degree of MB control. Third, the new methods will be assessed on their ability to compare the MB control between individuals and between the stages of a task. Finally, the new methods will be applied to compare between participants in a task believed to enhance the reliance on the planning strategy.

2 Methods

The experimental task

The task was an adaption of a two-stage paradigm designed by (Kool et al. 2016). The novel paradigm developed by Kool et al. (2018) allowed them to investigate the allocation of cognitive effort as a function of task complexity. The task is a multi-stage decision-making task with two levels of difficulty. The difference of difficulty levels is ensured by a different depth of the decision tree. During a session, high-effort or low-effort trials are presented randomly. The low effort task consists of a middle stage with one of two possible states presenting three spaceships each, and a terminal stage with one of three possible states presenting an alien. The high effort task consists of a top stage with one state presenting two space stations out of three, a middle stage with one of three possible states presenting two spaceships each, and a terminal stage, which is the same as the low effort task. At each trial, the choice of a space station or a spaceship leads to the subsequent stage following permanent deterministic transitions (e.g. the orange spaceship always leads to the yellow alien). The objects and transitions remain the same along the trials. The stimuli of the middle stage (spaceships) and the terminal stage (space-stations) are also the same for the high-effort and the low-effort and allows then to transpose knowledge from one trial to another. The objective of the task is to maximize the rewards given by the alien at the terminal state. The reward given by each alien drifts slowly following a Gaussian random walk ($\sigma = 2$), with bounds at 0 and 9. For the rest of the paper, the low effort middle stage, the high effort top stage and middle stage will be referred respectively as "low", "high0" and "high1".

Importantly, the structure of this paradigm makes it suitable to distinguish between choices made by a model-based or a model-free agent. Let's consider an initial low effort trial. The participant chooses one spaceship that leads to an alien and its associated reward. On the second trial, high effort or low effort, the initial state possesses different spaceships than the one previously selected. On the one hand, a model-free agent will make a totally exploratory choice as he never faced any of the spaceships. On the other hand, a model-based agent will plan to go to the same alien by selecting the right spaceship according to its knowledge of the task structure.

Model-free

The MF agent is built as a SARSA(λ) temporal difference learning algorithm (Rummery & Niranjan 1994). The Q value of the chosen action at time t is updated following:

$$Q_{MF}(s_i, a_i) = Q_{MF}(s_i, a_i) + \sum_{j=i}^{n} lr * \delta_{s_i,a_j} * \lambda^{n-j} \tag{1}$$

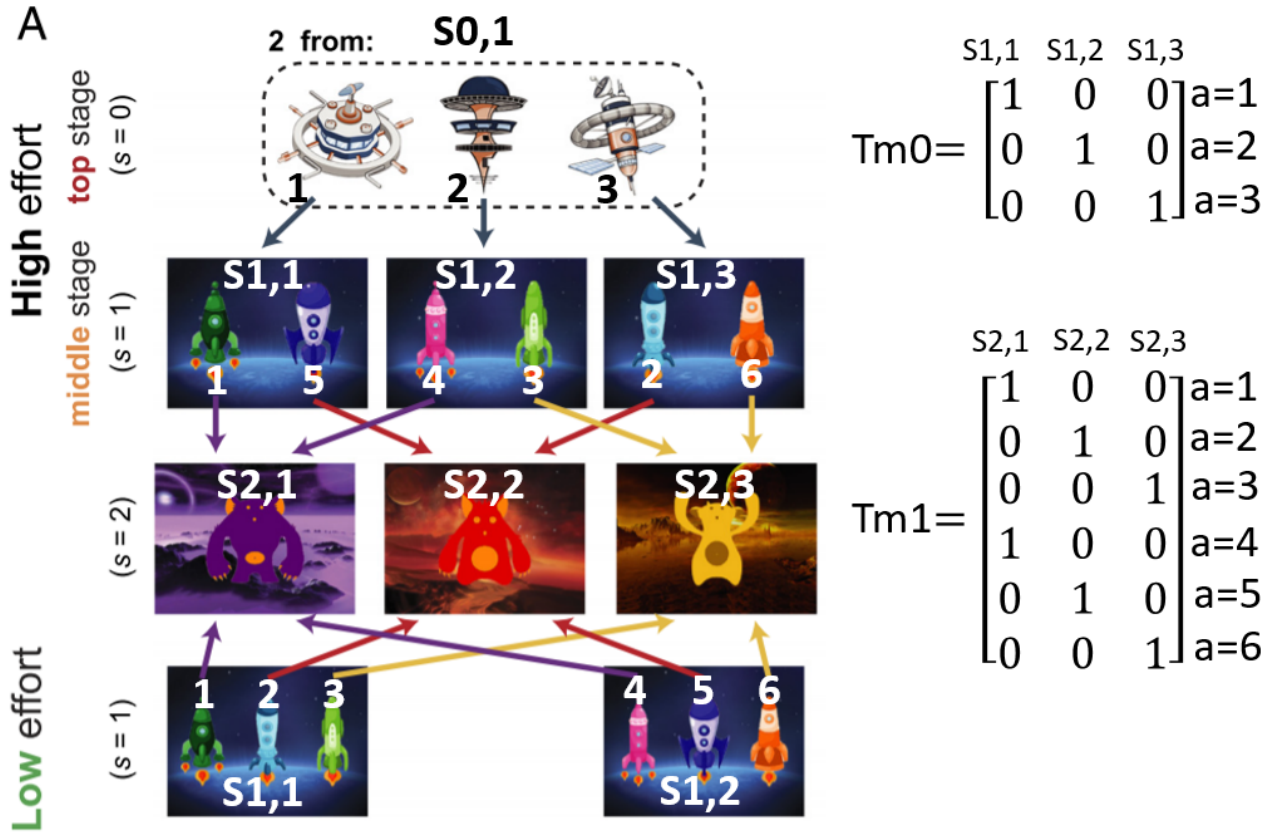

Figure 1: The multistage decision making task. $Tm0$ gives the transitions from the high-effort top stage to the high-effort middle stage. $Tm1$ gives the transitions from the low-effort or the high-effort middle stages to the terminal stage. The given transition matrices and the stimuli per stage are given as an example but vary across participants. Adapted from (Kool et al. 2018).

where n is the depth of the decision tree, lr the learning rate and λ the eligibility trace decay. The reward prediction error is defined as:

$$\delta_{s_i,a_j} = r(s_i) + Q_{MF}(s_{i+1}, a_{i+1}) - Q_{MF}(s_i, a_i) \tag{2}$$

Model-based

The MB agent computes the Q value of a state-action by inspecting the transition map to find which state at each stage leads to the maximum reward. This requires that the agent knows the transition matrices between each stage. At the terminal stage, MB and MF Q values are equivalent as there is no transition to an upper stage. Applying the MB agent to our task would give for the terminal stage:

$$Q_{MB}(s2) = Q_{MF}(s2) \tag{3}$$

We obtain the Q values of the middle stage and the top stage by multiplying the transition matrix Tm with the Q value of the terminal stage. The stimulus $stims_1$ gives respectively the indexes of the two spaceships or the three spaceships in presence at the state s_1 respectively for the high-effort or the low-effort task:

$$Q_{MB}(s_1) = Tm1(stims_1) * Q_{MB}(s_2) \tag{4}$$

In the case of an high effort trial, the agent is required to compute the Q values of the top stage. Being at the second level of the decision tree, two steps are needed. First we compute

the Q_{MB} values of the three states of the high-effort middle stage.

$$Q_{MB}(s_{1,i}) = Tm1(stims_{1,i}) * Q_{MB}(s_2); i \in 1, 2, 3 \tag{5}$$

Second, we select the maximum middle stage Q values per state. Then the top transition matrix gives which space-station to select to reach this optimal middle state. The stimulus $stims_0$ gives the indexes of the two space-stations in presence at the state s_0.

$$Q_{MB}(s_0) = Tm0(stims_0) * max_{a \in A_{s_1}}(Q_{MB}(s_1, a)) \tag{6}$$

Dual-system RL Model

The dual system is an hybrid agent of the MB and MF models (MBMF model). The Q values for each stage are mixed according to a weight $0 < w < 1$ (Daw et al. 2011):

$$Q = w * Q_{MB} + (1 - w) * Q_{MF} \tag{7}$$

The decision rule relies on the soft-max activation function to transform the Q values in a probability to select an action. An action is taken randomly following this probability distribution.

$$Ps(a_i) = \frac{\exp \beta * Q(a_i)}{\sum_{a \in A} \exp \beta * Q(a)} \tag{8}$$

where A is the set of available actions at the current state and β the inverse temperature defining the exploration. A β close to 0 means total exploration while a β tending to infinity means no exploration.
The weights w can be assumed as different for each stage. In this case we will set w_{low}, w_{high0} and w_{high1} the respective weights of the middle stage low effort, the top stage high effort and the middle stage high effort. Similarly, assuming different exploration rates conducts to define β_{low}, β_{high0} and β_{high1}. We will call *exhaustive* an agent with a β parameter per stage. Conversely, an agent possessing the same β for all the stages will be called *simple* (see Table 1). The codes implementing these methods are available on (Paques 2021).

Model	Free parameters
simple MB	$[\beta, lr]$
simple MBMF	$[\beta, w_{low}, w_{high0}, w_{high1}, lr, \lambda]$
exhaustive MB	$[\beta_{low}, \beta_{high0}, \beta_{high1}, lr]$
exhaustive MBMF	$[\beta_{low}, \beta_{high0}, \beta_{high1}, lr, \lambda, w_{low}, w_{high0}, w_{high1}]$

Table 1: Nomenclature of the RL models

Model Fitting Procedure

Each of the aforementioned models has several free parameters that need to be fitted on participants' data. The model fitting procedure made use of the $mfit$ toolbox (Gershman 2016) to fit the free parameters of the different RL models. The aforementioned tool allows one to use two different methods: the Maximum A Posteriori (MAP) and the Hierarchical Bayesian Maximum A Posteriori (HBMAP) estimations. Both procedures are Bayesian-based approaches to estimate the model parameters that best describe a set of data. The MAP estimates the posterior probability of a parameter α given a noisy observation $\hat{f}$: $p(\alpha|\hat{f}) = p(\hat{f}|\alpha) * p(\alpha)$. On the other hand, the HBMAP is a hierarchical method. The estimation model is split into sub-models and

the uncertainties are propagated from one to another. Now, a nuisance parameter θ is added and the posterior probability is estimated depending on a noise term, a theory term, and priors: $p(\theta, \alpha, |\hat{f}) = p(\hat{f}|\theta) * p(\theta|\alpha) * p(\alpha)$.

In our case, we search to maximize the log likelihood of a RL model by fitting its parameters. The function modeling the agent takes the participant data as an input and outputs the sum of the log likelihood of each stage (e.g. exhaustive MB agent Section 5, 1). The log likelihood is defined as the log of the sum of the probability to take an action at time t.

$$LL = \sum_{t=0}^{N} log \frac{\exp \beta * Q(s_t, a_t)}{\sum_{a \in A_t} \exp \beta * Q(s_t, a)} \tag{9}$$

According to empirical results, the inverse temperature β prior was set as a Gamma distribution $\Gamma(4.82, 0.88)$ between 0 and 20 (Gershman 2016). The priors of the other parameters had a uniform distribution between 0 and 1. The fitting procedure was performed twice using different random initialization of the parameters. Each participant's data was fitted separately allowing to perform statistical analysis beyond a population of participants. To feed the RL model, the reward was simply normalized by dividing by 9. To remedy extensive computational time of execution, the fitting method will be the MAP, instead the contrary is explicitly specified, .

Data production

An appealing way to assess the fitting procedure is to create data from "machine participants". If one can know the exact strategy followed by the subjects, then one has a reference to compare the estimated strategy. The exhaustive MBMF agent was used to generate the data (see Section 5, Algorithm 2). The agent evolves in the same environment as the Kool et al. experiment and takes its own decisions based on a mixed model-based and model-free strategy. The values of the eight parameters are taken as an input and the algorithm outputs the different states, actions, and rewards obtained during the trials. The resulting data are suitable to fit any aforementioned model using the fitting procedure.

In order to know what sets of parameters were relevant, we investigated the evolution of the log likelihood as a function of the weights and the exploration rates (see Figure 2a)). Logically, the log likelihood increases with the inverse temperature β. The less exploration, the more likely the fitted agent reproduces the same decisions as the simulated agent. It appears that for large values of βs ($\beta \geq 15$) the log likelihood reaches a plateau. Despite a supposedly more deterministic behaviour, the fitted model doesn't improve its prediction of the simulated model. This convergence of the estimation performances implies that after a threshold, the agent explores - or more accurately exploits - to the same extent whatever the β values. This set of β should be avoided as it is preferable that two degrees of exploration are distinguishable to study the effect of MB control. To investigate the behaviour of the MB likelihood, we fitted a MB agent using simulation data with different values of β and w. The 3D plot of the log likelihoods shown that for the same value of β and w the log likelihoods are different for the three stages (see Figure 2b)).

In Kool et al., the MB control is associated to the weights of the MBMF agent. However, it seems acceptable to assume that the reliance on a control strategy would equally depends on the exploration. Thereby, in the case of data produced by the exhaustive MBMF agent, we define the degree of MB control as $degree_{MB} = \beta * w$ as it has been done in (Milena Rmus et al. 2019).

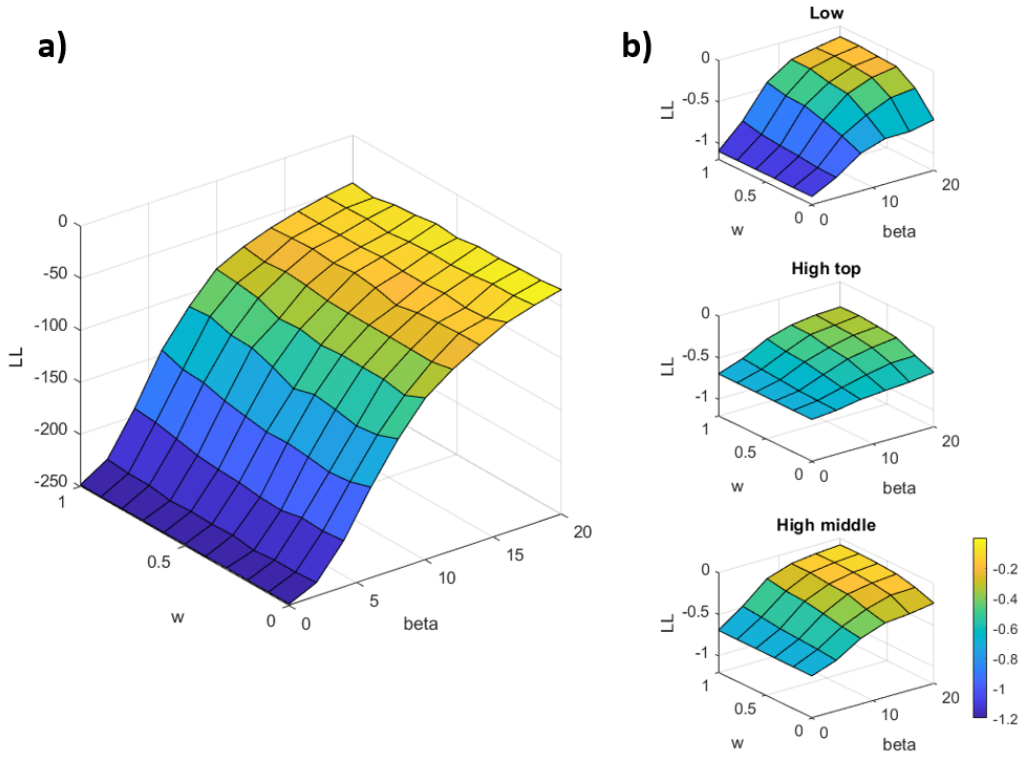

Figure 2: a): Log likelihood map of an exhaustive MBMF agent. **b)**: Normalized Log likelihood map of an exhaustive MB agent for the three stages. Top: Low-effort, Middle: High-effort top, Bottom: High-effort middle.

3 Results

The dual system

Kool et al. results

The objective of the experiment designed by Kool et al. (2018) was to verify if participants will prefer a less accurate model-free system when planning complexity increases. Data from 101 participants were collected for the aforementioned experiment. The parameters of a simple MBMF agent were estimated using the MAP fitting procedure. The planning complexity had a statistically significant effect: participants presented a smaller MBMF weight for the high effort top trial than for the low effort trial (mean $w_{high0} = 0.50$, mean $w_{low} = 0.70$, $t(97) = 4.46$, $p < 0.001$, $d = 0.45$). However, no difference was found between the fitted weights of the low effort and the high-effort middle stages ($t(97) < 1$). Kool et al. also found a smaller reward rate for the high-effort consistent with the use of a less accurate strategy.

The presented results offered a coherent illustration of the trade-off between the task complexity and the degree of MB control. However, by using the MBMF simple agent, Kool et al. made the strong assumption that exploration rate was uniform across the stages. Conscious of this concern, they also fitted an exhaustive MBMF agent and found out a significant difference of inverse temperature between the low effort and the high effort top stages and between the high effort top and high effort middle stages. However, as the weight parameters still presented significant results respecting the complexity effect, the hypothesis of different βs was considered unnecessary.

Noisy estimation

Here came our intuition that a deeper investigation of the fitting procedure was needed. Indeed, the exploration rate determines the randomness of the choices and must therefore have an influence on the estimated degree of MB control. A correlation study on participants' data showed a positive correlation for all β and w and a significant positive correlation between w_{low} and β_{low} ($r = 0.25$; $p = 9.8e-3$) and between w_{high1} and β_{low} ($r = 0.28$; $p = 4.8e-3$). The smaller the inverse temperature, the more exploration and thus the more noise in the decisions taken. In the fitting procedure of Kool et al. the weights were assumed to follow a uniform distribution $U(0,1)$ (see Section 2, Fitting procedure). In case of very noisy decision data, it is predictable that the weights estimations tends toward 0.5, the mean of the interval [0,1]. In the current task, the weights were estimated over 0.5 whatever the stages. Therefore, a decrease of β engenders a decrease of w and explains the positive correlation. To investigate the ability of the fitting procedure to account for the strategy of a simulated model we created our own data (see Section 2, Data production). The fitted parameters of two RL models are presented in Table 2. Both the exhaustive MBMF and to a larger extent the simple MBMF failed to recover the simulation parameters. Both models estimated a smaller weight with high significance for the high effort top stage (low/high0: p_{simple}=1.1e-11, $p_{exaust.}$=3.2e-3; high1/high0: p_{simple}=3.4e-09, $p_{exaust.}$=2.1e-2). The chosen simulation parameters might not be biologically plausible as one can wonder why a participant would rely on a goal directed strategy (high w) while exploring massively (low β). Rather than saying the presented fitting method can't describe Human behaviour our point was to caution on the interpretation of the fitted parameters which could in theory be erroneous and in practice very likely be overestimated.

Fitting procedure						
Model	β_{low}	β_{high0}	β_{high1}	w_{low}	w_{high0}	w_{high1}
exhaustive MBMF (Simul.)	5	2	5	0.7	0.7	0.7
simple MBMF (Fitted)	4.44	4.44	4.44	0.78	0.53	0.75
exhaustive MBMF (Fitted)	4.79	2.75	4.81	0.71	0.60	0.70

Table 2: Mean simulation and fitted parameters for 98 participants (human or machine). Simulation and fitted model: exhaustive MBMF agent.

The fitting procedure highly depends on the number of data available. We expect the error on the β and w estimates to decrease for a higher number of trials. The evolution of the mean absolute error of the two parameters is observable on figure 3. More importantly, the t-test p values of the weights are reported and shows that after 1200 trials, the weights of the participants don't show a significant difference across the stages (p>0.3). This threshold suggests that for a large enough number of trials the fitting procedure gives satisfying results. However, two reservations must be given: first the data to be fitted were produced by a pure MBMF model and second such a high number of trials might not be feasible with human participants.

The dual system is too restrictive

The choice of a hybrid model to describe human behaviour is in itself questionable. Whilst the MBMF agent offers a consistent behavioural description, it is more likely that humans rely on more complex strategies than a simple model-based model-free dichotomy. Indeed, the two RL models present no temporal variability in the parameters. Humans also experience forgetting or fatigue, mechanisms that are omitted in these simple models. In addition, a participant misconceiving the task isn't rare. In a similar two-step reward decision task, subjects were shown to follow strategies relying on irrelevant features as the location of stimuli or the keys

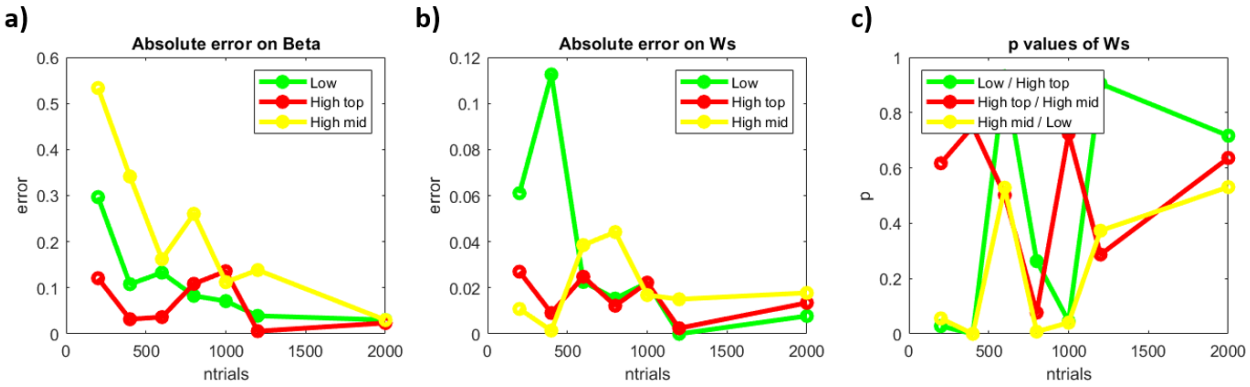

Figure 3: Error on parameters estimation versus number of trials performed by the participants. **a)** Mean absolute error on βs versus the number of trials. **b)** Mean absolute error on ws versus the number of trials. **c)** Mean p value of the null hypothesis between w_i and w_j versus the number of trials. Index i and j represents alternatively the low-effort, the high-effort top and the high-effort middle stages. The data were obtained by simulating an exhaustive MBMF agent with 98 machine participants and parameters : [β_{low}=2, β_{high0}=4 β_{high1}=6, w_{low}=w_{high0}=w_{high1}=0.7].

used for selection, instead of the stimuli semantics (Shahar et al. 2019, Feher da Silva & Hare 2020). If the strategies used rely on different goals and transitions systems, the MBMF model is inadequate.
A new approach would be to measure how likely participants data are described by the control strategy instead of relying on the fitted parameters of a hybrid model.

A new approach

The MB agent log likelihood

Our objective is to assess how well the participants' data of each stage is described by a MB agent. Using the MB agent likelihood of the different stages would in theory allow us to compare between participants. In figure 2b), the log likelihood of the fitted exhaustive MB agent revealed to increase with both the inverse temperature and the MB weight. Thus, a difference of MB likelihoods between two participants could inform on both their respective exploration and degree of MB control. When two groups of machine participants shared the same inverse temperatures but had different weights across the stages, the differences of their log likelihood respected the hierarchy of the weights (see figure 4a): mean(Δw)=[0.2, 0, -0.2]; mean(ΔLL)=[5.7e-2, -3.8e-2, -1.6e-2]). Similarly, when two groups shared the same weights and different inverse temperatures across the stages, the hierarchy of the inverse temperatures is respected (see figure 4b): mean($\Delta\beta$)=[2, 0, -2]; mean(ΔLL)=[1.3e-1, 1.8e-3, -8.3e-2]). This results indicates the MB log likelihood could be used to compare the degree of MB control between participants.

To study the complexity effect, another interesting information was to know what degree of MB control is associated with each stage for the same participant. However, one can compare likelihoods only if they are obtained by the same model on similar tasks. Here the tasks were rather different considering the number of possible actions and their position in the decision tree. Moreover, the design of the paradigm induced more randomness for the high-effort top stage than the two other stages. For example, considering the high effort top stage, in a third of the cases the two available space-stations gives the possibility to reach the alien with the maximum reward. Therefore, the two associated Q_{MB} values are equal. For this particular case, a MB agent and a random agent are as good. Considering this, one should be careful when

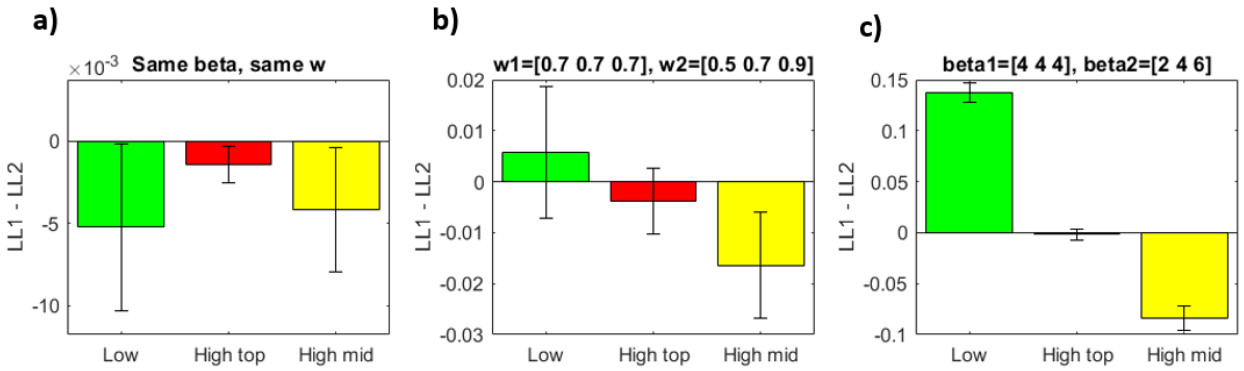

Figure 4: Difference of the MB log likelihoods between two groups of machine participants. Error bar indicates within-subject SEM. **a)** Same weights and same inverse temperatures. Group1 and Group2: w_{low}=w_{high0}=w_{high1}=0.7; β_{low}=β_{high0}=β_{high1}=4. **b)** Same inverse temperatures and different weights. Group1: w_{low}=w_{high0}=w_{high1}=0.7. Group2: w_{low}=0.5, w_{high0}=0.7, w_{high1}=0.9 **c)** Same weights and different inverse temperatures. Group1: β_{low}=β_{high0}=β_{high1}=4. Group2: β_{low}=2, β_{high0}=4, β_{high1}=6.

comparing the MB likelihoods across the stages. Similarly, the log likelihood map of the fitted MB agent revealed that for the same value of β and w the log likelihoods of each stages are different (see Figure 2b)).

This observation motivates the introduction of an extra agent to compare the likelihood with and thus potentially enabling the comparison across stages. We introduced four different metrics with the hope they embody the degree of MB control. To assess these metrics we produced data from "machine participants" using an exhaustive MBMF agent performing the same multistage task (see Section 2). These metrics should be able to account for both the exploration rates and the weights of the simulated agent. We must confess the definition of the metrics has more to do with simple logic than real theory as the log likelihood depends on the complexity of the task, the structure of the paradigm, and many other factors.

*Metrics 1 : Difference with a Random agent
The metrics 1 is defined as:

$$metrics_1 = LL_{MB} - LL_{Random} \tag{10}$$

where LL_{MB} is the log likelihood of a MB model which parameters have been fitted on the participants' data and LL_{Random} the log likelihood of a random model. The idea of this metric is to measure how far the participants' strategy is from the random strategy. Considering that the MB strategy is optimal and the Random strategy sub-optimal, one can expect that the difference of their likelihoods would be positive and maximal for a pure MB participant. The random agent simply selects randomly an action according to an uniform probability distribution.

*Metrics 2 : Difference with an optimal MB agent
The metrics 2 is defined as:

$$metrics_2 = LL_{MB} - LL_{MB,optimal} \tag{11}$$

where the $LL_{MB,optimal}$ is the log likelihood of an optimal agent evolving in the same environment as the participant (same rewards, same initial states). Using the reverse logic as metrics 1, we decide here to compare the participants log likelihood with an optimal agent. The optimal agent is defined as an MB agent selecting always the best options. This agent is omniscient, as it knows the environment structure as well as what rewards are allocated in real time. Therefore it never explores nor needs to learn. One would expect the more degree of MB control, the smaller difference of likelihoods.

*Metrics 3 : Difference with a machine MB agent
The metrics 3 is defined as:

$$metrics_3 = LL_{MB} - LL_{MB,machine} \tag{12}$$

where $LL_{MB,machine}$ is the log likelihood of an independent MB agent. The Machine agent is a MB agent evolving in the same environment as the participant but taking its own decisions. Its parameters are the same as the parameters of the MB agent fitted on the participants' data. If the participant relies purely on the MB agent the difference of likelihoods should be minimum. We condition this statement to a limited exploration.

*Metrics 4 : Ratio of the difference
The metrics 4 is defined as:

$$metrics_4 = \frac{LL_{MB} - LL_{Random}}{LL_{MB,optimal} - LL_{Random}} \tag{13}$$

We introduced a last metrics combining the log likelihoods of the previous agents. Still considering that the Random agent is the most distant agent from the MB agent, we expect the ratio to be close to unity when the participant follows a control strategy and close to zero otherwise.

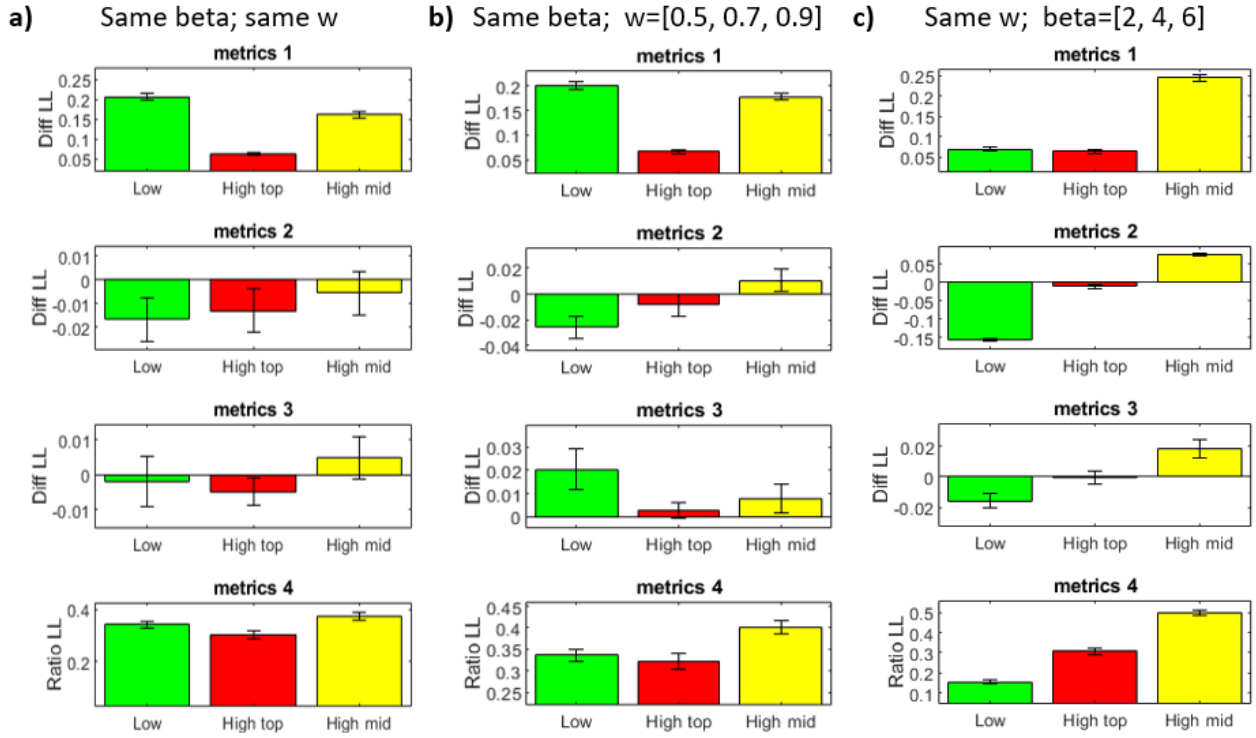

Figure 5: Mean value of the metrics across the stages for 98 machine participants. Error bars indicate within-subject SEM. When indicated same w, w_{low}=w_{high0}=w_{high1}=0.7. When indicated same β, β_{low}=β_{high0}=β_{high1}=4.

A first step in the evaluation of the metrics is to verify if they offer a consistent behaviour regarding the strategy of a simulated agent. We produced data by simulating the same agent but with different parameters as described in Section 2, Data production. A good metrics will be able to account for an increase of exploration and / or an increase of the MB weights across the stages. The mean values of the fitted parameters are presented in figure 5. On Figure 5a), the metrics 1 and 4 don't report the invariance of exploration nor the degree of MB control across the stages. On the other hand, metrics 2 and 3 present no significant differences across

the stages (metrics 2: p >0.20, metrics 3: p >0.28). On Figure 5b), participants' data are produced with the same βs but different ws (w_{low}=0.5, w_{high0}=0.7, w_{high1}=0.9). One would expect a positive correlation between the simulation weights and the metrics values. This time, the hierarchy was only respected for metrics 2 with acceptable significance (mean($metrics_2$) =[-2.5e-2, -8.3e-3, 2.0e-2], p <0.01). Finally, on figure 5c), the simulation weights are the same but the exploration is different (β_{low}=2, β_{high0}=4, β_{high1}=6). For these data, every metrics but $metrics_1$ shown a correct macroscopic behaviour with good significance ($p < 10^{-3}$).

The same analysis was performed using the HBMAP fitting procedure. Using a Hierarchical model was thought to decrease the noisy estimation of the parameters. The same general behaviour was observed for every metrics. However, no significant improvement was brought by this fitting method. For the data obtained by simulation with the same weights and inverse temperatures the difference between the values obtained across the stages were slightly more important (metrics 2: p >0.05, metrics 3: p >0.2).

Validity and Reliability

To validate a measurement method, one need to guarantee the validity, the reliability and the accuracy of the measures. Note that in our case, the accuracy can't be assessed as our method doesn't have the ambition to give the degree of MB control but to allow a comparison of the strategy used between participants or between stages. Therefore, the validation of the metrics concerns only the validity and the reliability. The validity can be obtained by making a correlation study between the measure and an outside valid criterion. In our particular case, an outside valid criterion could be the model-based control evaluated using the parameters from the simulated agent. Three metrics will be compared: 1) the log likelihood of a fitted exhaustive MB agent LL, 2) the $metrics_2$ and 3) the product of the weight and the inverse temperature of a fitted exhaustive MBMF agent $(\beta w)_{fitted}$.
The validity study was carried out on 30 groups of 98 machine participants whose data were produced as described in Section 2/Data production. Each group of data were generated using different values of β and w for each stage taken randomly respectively from U(0,16) and U(0,1). The obtained results shown a positive correlation between the degree of MB control and LL of the same stages ($r > 0.73$; $p < 1e-6$) (see Table 3). Very similar results were obtained for $metrics_2$ ($r > 0.74$; $p < 1e-6$). The metrics $(\beta w)_{fitted}$ had weaker correlations specially for the high-effort middle stage and top stage ($r > 0.55$, $p < 1e-2$).

The reliability addresses if the measures are reproducible for a same stimulus. Using the same previous data, an exhaustive MB agent was fitted twice to produce two sets of metrics for the three stages. Then, the t-test hypothesis was measured between the first and second fitted sets of metrics. Finally, the mean values of the statistics variables across the groups were calculated for each stage. The reliability study shown satisfying repeatability performances for $(\beta * w)_{fitted}$ ($p > 0.35$, $t < 0.3$, $d < 3e-2$) and slightly better ones for $metrics_2$ and the MB log likelihoods ($p > 0.40$, $t < 0.3$, $d < 3e-2$).

The inter-stages reliability of the metrics was also verified when the simulation agent had the same parameters across the stages. The metrics are expected to give the same value as the degree of MB control is the same for the three stages. Thirty simulations were run for β in [0, 15] and w in [0, 1] and data of each simulated group were fitted. As expected from the log likelihood maps (see Figure 2b), the inter-stages difference of the MB likelihood was very strong ($p < 1e-2$, $t > 20$, $d \leq 13$). The difference between two stages for a same group also proved to be noticeable for $metrics_2$ but to a lesser extent. Specifically, the difference for the low-effort and the high-effort middle stages proved to be remarkably smaller (LL: $p = 1e-19$, $t = 118$, $d = 11$; $metrics_2$: $p = 6e-2$, $t = 5$, $d = 0.5$). The $(\beta * w)_{fitted}$ metrics presented the better repeatability performances across the stages without however showing no significant

Validity			
correlation test	$(\beta w)_{low}$	$(\beta w)_{high0}$	$(\beta w)_{high1}$
LL_{low}	r=0.84; p=3.1e-09	r=0.37; p=4.2e-2	r=-0.44; p=1.4e-2
LL_{high0}	r=0.08; p=0.63	r=0.73; p=3.1e-06	r=-0.35; p=5.7e-2
LL_{high1}	r=-0.40; p=2.5e-2	r=-0.36; p=4.8e-2	r=0.85; p=1.2e-09
$metrics_{2,low}$	r=0.84; p=3.3e-9	r=0.37; p=4.1e-2	r=-0.44; p=1.4e-2
$metrics_{2,high0}$	r=0.09; p=0.63	r=0.74; p=2.7e-6	r=-0.35; p=5.6e-2
$metrics_{2,high1}$	r=-0.40; p=2.5e-2	r=-0.36; p=4.7e-2	r=0.85 ;p=1.1e-9
$(\beta w)_{fitted,low}$	r=0.73; p=4.6e-06	r=0.31; p=8.8e-2	r=-0.23; p=0.21
$(\beta w)_{fitted,high0}$	r=0.14; p=0.43	r=0.55; p=1.5e-3	r=-0.29; p=0.11
$(\beta w)_{fitted,high1}$	r=-0.23; p=0.20	r=-0.40; p=2.6e-2	r=0.55; p=1.2e-3
Reliability			
t-test	low	high0	high1
LL	p=0.41 t=6.9e-2 d=7.0e-3	p=0.45 t=4.3e-2 d=4.4e-3	p=0.40 t=0.22 d=2.2e-2
$metrics_2$	p=0.41 t=6.9e-2 d=7.0e-3	p=0.45 t=4.3e-2 d=4.4e-3	p=0.40 t=0.22 d=2.2e-2
$(\beta w)_{fitted}$	p=0.41 t=0.13 d=1.3e-2	p=0.35 t=0.16 d=1.6e-2	p=0.42 t=0.21 d=2.1e-2

Table 3: Validity and Reliability of the metrics. **Validity:** Correlation between the mean of the metrics for a group and the degree of MB control. **Reliability:** Mean correlation between two sets of metrics of a same stage within a group. Random β and w for the 3 stages of a group. 30 groups, 98 participants per group, 200 trials per participant.

differences (low and high1 difference: $t > 1.7$; high1 and low difference: $t > 1.1$).

Inter-stages Reliability			
t-test	low and high0	high0 and high1	high1 and low
LL	p=6.7e-2, t=130, d=13	p=9.9e-2, t=22, d=-2	p=1.7e-19, t=118, d=11
$metrics_2$	p=5.8e-2, t=2.875, d=0.29	p=6.4e-2, t=2.9, d=0.29	p=6.1e-2, t=-5.5, d=-0.56
$(\beta w)_{fitted}$	p=0.20, t=1.7, d=0.17	p=0.41, t=0.46, d=4.6e-2	p=0.37, t=1.1, d=0.10

Table 4: Inter-stage reliability of the metrics. Mean correlation between the metrics of two stages within a group. Same β and w for the 3 stages of a group. 30 groups, 98 participants per group, 200 trials per participant.

Comparing between human participants

In the same paper, Kool et al. updated the multistage task to enhance the use of a goal-directed strategy ((Kool et al. 2018), Experiment 2). The model-based strategy is believed to be preferred when cognitive resources are available. The task structure and the actions remained the same but longer response time was allowed (10 sec instead of 2 sec) and participants were trained more extensively to memorize the transition map. In addition, probe trials were added, for which a massive reward was announced to the participant. The probe trials were believed to favour the motivation to plan more carefully. With regard to these changes, one would expect a higher degree of MB control for participants on the Experiment 2 than on the Experiment 1. The three metrics were calculated by fitting their respective model on the data of Human participants from Experiment 1 and Experiment 2. We recall LL and $metrics2$ are obtained from an exhaustive MB model while $(\beta w)_{fitted}$ is obtained from the parameters of an exhaustive MBMF model. The plot of the mean of the metrics is presented on Figure 6. For each metrics and for each stage a smaller MB control was measured. However, the difference between the two experiments was significantly bigger for the LL and $metrics_2$ than for $(\beta w)_{fitted}$ for the

low-effort stage (LL and $metrics_2$: $t > 1.7$, $p < 7.5e-2$; $(\beta w)_{fitted}$: $t = 1.4$, $p = 1.5e-1$) and the high-effort middle stage (LL and $metrics_2$: $t > 1.9$, $p < 5.0e-2$; $(\beta w)_{fitted}$: $t = 1.3$, $p = 1.7e-1$). The difference between the metrics were comparably significant on the high-effort top stage ($t > 4$, $p < 3e-4$).

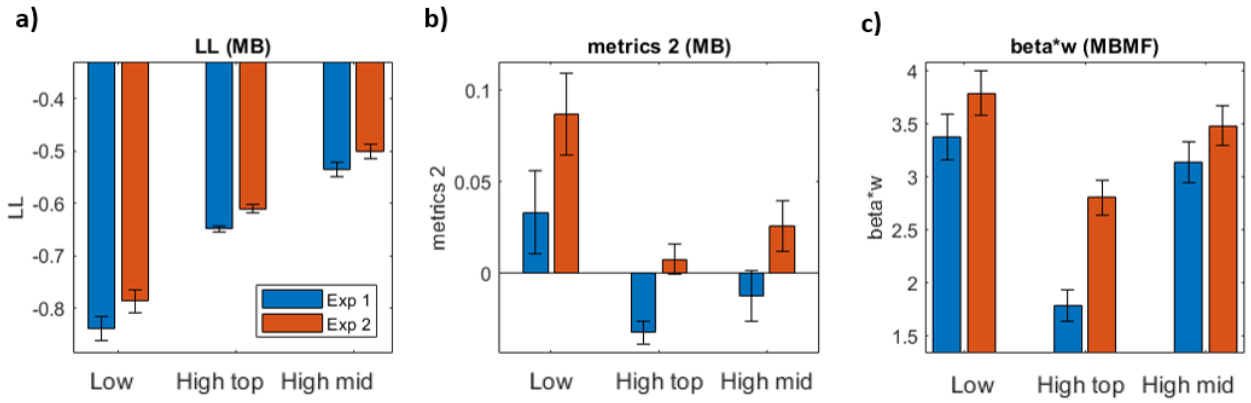

Figure 6: Comparing the degree of MB control between Experiment 1 and Experiment 2. Error bars indicate within-subject SEM. **a)** MB log likelihood LL **b)** $metrics_2$ **c)** $(\beta w)_{fitted}$

4 Discussion

A common RL model to describe Human decision is a hybrid model mixing a model-free and a model-based strategy (MBMF agent). Relying on an internal map of the task, the model-based is more accurate but also more computationally expensive. The model-free performs fast but inaccurate decisions by storing the expected values of the previous experiences. Whilst the model-based strategy benefits from multiple neural and cognitive evidences (Gillan et al. 2015, Doll et al. 2015, Daw et al. 2011), the existence of a model-free strategy is more controversial. Conversely to the model-based, no studies were able to identify an internal brain representation of the model-free (Miller et al. 2017). Moreover, the behavioral evidences of the MF are often weak and gives rise to alternative explanations (Miller et al. 2019, Feher da Silva & Hare 2020). The original idea of this paper is to measure the degree of MB control without relying on the weights of a dual model but by using the likelihood of a MB agent. To investigate the complexity effect on the mental effort allocation, Kool et al. designed a multistage decision-making task and analyzed the participants' responses using the hybrid MBMF model (Kool et al. 2018). Studying the complexity effect required comparing the degree of MB control across the stages. In the current paper, we proposed to exploit their data and task as a base to search for new ways to measure the degree of MB control between participants and across the stages. A first control of Kool et al. fitting procedure revealed that the weights estimation suffered from the noise inherent to the exploration. The estimation can be improved by increasing the number of free parameters of the MBMF agent, more particularly by having an inverse temperature parameter β per stage (see Table 2). Another source of improvement lies in a higher number of trials, what is realisable for machine participants but can turn out to be more demanding for Humans (see Figure 3).

The use of the MB likelihood revealed to correctly estimates the difference of degree of MB control between two population of participants (see Figure 4). The comparison of the two groups gave consistent results when the simulation βs and ws were different across the stages. However, to compare the MB control between stages, the log likelihood revealed to be a poor metrics. Indeed, the log likelihood doesn't allow a fair comparison for tasks with different number of options and randomness. Therefore a new metrics was needed implying a

comparison with an extra agent. We defined four of them using a random agent, an optimal MB agent and a competitor MB agent. The general behaviour of the metrics isn't a precise way to assess their validity but it has the merit to highlight the ones to avoid. Data were produced by a hybrid MBMF model whose parameters were set to obtain different exploratory and model-based configurations. The analysis of the metrics applied on these data revealed a correct behaviour for the metrics defined as the difference between the log likelihood of a fitted exhaustive MB agent and the log likelihood of an optimal MB agent ($metrics_2$). The metrics was able to account for an invariance of MB control across the stages as well as assessing a difference of the inverse temperatures and the MBMF weights of the simulated agent (see figure 5). To perform a more rigorous validation of the proposed methods, two qualities of the measurement were inspected: the validity and the reliability. The log likelihood of the MB agent and $metrics_2$ proved to be reliable and valid to compare the degree of MB control between participants. The product of the weights and inverse temperatures of a fitted exhaustive MBMF agent $(\beta * w)_{fitted}$ also shown good reliability results but a moderate validity for the high-effort top and middle stages (see Table 3). Comparing the MB control across the stages revealed to be a more complex feature to measure. None of the three metrics gave satisfying reliability results (see Table 4). Despite encouraging reliability performances for $metrics_2$, the design of a metrics enabling for a fair comparison across the stages was a failure. The reliance on the hybrid model parameters proved to be inefficient as well. Finally, an application of our metrics was made to compare Human participants. The comparison opposed the data from the presented multistage task and the data obtain from an extended form of the multistage task, designed to enhance the use of MB control (see (Kool et al. 2018) Experiment 2). Using Human data, the MB log likelihood and $metrics_2$ confirmed their advantage over $(\beta * w)_{fitted}$ to compare the MB control between participants.

With regard to the two investigated measures, we advise against the use of the hybrid model parameters to account for the MB control in favour of a method using the MB likelihood. We show better performances for the MB likelihood on comparing between participants on the same task. Despite promising results for $metrics_2$, we claim that none of the presented methods are able to compare the degree of MB control across stages for this multistage task.

A major limitation of this work lies in the means used to assess the proposed methods. Indeed, the quality of the metrics were controlled using data produced by an exhaustive MBMF model. Two problems came out of this choice. First, the $(\beta * w)_{fitted}$ method was advantaged because the fitted model was the same as the simulation model. Thus, the poor validity performance of this last method should especially question its use. Second, our results are only confirmed for an exhaustive MBMF agent and several studies suggest the mixed model doesn't embody human decision-making (Feher da Silva & Hare 2020, Miller et al. 2019). Therefore, the quality control of our method isn't guaranteed on Human data. Further testing should be done using Human data from an experiment specially conceived to enhance the MB strategy of one group compared to another. A shorter response time allocated, the requirement to perform simultaneously a disruptive task and the absence of prior knowledge of the structure could be the features of the disfavoured group ((Otto et al. 2013), Exp. 2; (Kool et al. 2018), Exp. 2). Another appealing application of the likelihood method could be the diagnosis of people suffering obsessive-compulsive disorders (OCD). OCD patients are characterized by a lesser ability to plan (Gruner et al. 2016, Bey et al. 2018). Considering our good results to compare participants, we have hope our method would more significantly discriminate OCD from healthy patients. To ensure a fair comparison between the fitted agents, another idea would be to use a different computational model than the MBMF. Instead of a dual model, the data could be produced by a quadruple model as suggested by (Miller et al. 2017). One model would account for planning (MB agent), one for perseveration, one for novelty preference and one for the bias toward a choice. A second questionable feature of our method is the absence of forgetting mechanism for

the RL model. The task structure comprehends numerous actions and transitions in addition to drifting rewards that the participant needs to memorize along the two hundred trials. The presence of a forgetting process appears a relevant update to the model. Using the SARSA(λ) algorithm as a base, a forgetting step could be added by regressing the Q values of the unselected actions toward a same scalar proportionally to a forgetting rate. This approach has shown to increase the likelihood of a fitted RL model in a very similar multistage task (Toyama et al. 2019).

References

Bey, K., Kaufmann, C., Lennertz, L., Riesel, A., Klawohn, J.and Heinzel, S., Grützmann, R., Kathmann, N. & Wagner, M. (2018), 'Impaired planning in patients with obsessive-compulsive disorder and unaffected first-degree relatives: Evidence for a cognitive endophenotype.', *Anxiety Disord.* .

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**, 1204–1215.

Doll, B., Duncan, K., Simon, D., Shohamy, D. & Daw, N. (2015), 'Model-based choices involve prospective neural activity.', *Nat Neurosci.* **5**, 767–72.

Feher da Silva, C. & Hare, T. A. (2020), 'Humans primarily use model-based inference in the two-stage task.', *Nat Hum Behav.* **10**, 1053–1066.

Gershman, S. J. (2016), 'Empirical priors for reinforcement learning models.', *Journal of Mathematical Psychology* **71**, 1–6.

Gillan, C., Kosinski, M., Whelan, R., Phelps, E. & Daw, N. (2016), 'Characterizing a psychiatric symptom dimension related to deficits in goal-directed control.', *Elife* .

Gillan, C. M., Otto, A. R., Phelps, E. A. & Daw, N. D. (2015), 'Model-based learning protects against forming habits.', *Cognitive, Affective Behavioral Neuroscience* **15**(3), 523–536.

Gruner, P., Anticevic, A., Lee, D. & Pittenger, C. (2016), 'Arbitration between action strategies in obsessive-compulsive disorder.', *The Neuroscientist : a review journal bringing neurobiology, neurology and psychiatry.* **22**(2), 188–198.

Huettel, S. A., Mack, P. B. & McCarthy, G. (2002), 'Perceiving patterns in random series: dynamic processing of sequence in prefrontal cortex.', *Nature Neuroscience* **14**.

Kahneman, D. (2003), 'A perspective on judgment and choice: Mapping bounded rationality.', *American Psychologist* **A**(58), 697–720.

Kool, W., Cushman, F. & S.J., G. (2016), 'When does model-based control pay off?', *PLoS Comput Biol.* **12**(8).

Kool, W., Gershman, S. J. & Cushman, F. A. (1985), 'Actions and habits: The development of behavioural autonomy.', *Philosophical Transactions of the Royal Society B: Biological Sciences* **308,**, 67–78.

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.

Milena Rmus, M., Ritz, H., Hunter, L., Bornstein, A. M. & Shenhav, A. (2019), 'Individual differences in model-based planning are linked to the ability to infer latent structure.', *BioRxiv* .

Miller, K., Botvinick & M. Brody, C. (2017), 'Dorsal hippocampus contributes to model-based planning.', *Nat Neurosci* **20**, 1269–1276.

Miller, K., Shenhav, A. & Ludvig, E. (2019), 'Habits without values.', *Psychol Rev.* **126**(2), 292–311.

Newell, B. R. & Schulze, C. (2000), 'Probability matching.', *Cognitive Illusions: Intriguing Phenomena in Judgement, Tinking and Memory, chap. 3, 50* **14**.

Otto, A. R., Gershman, S. J., Markman, A. B. & Daw, N. D. (2013), 'The curse of planning: Dissecting multiple reinforcement-learning systems by taxing the central executive.', *Psychological Science* .

Otto, A. R., Skatova, A., Madlon-Kay, S. & Daw, N. D. (2015), 'Cognitive control predicts use of model-based reinforcement learning.', *Journal of Cognitive Neuroscience* **27**(2), 319–333.

Paques, M. (2021), 'https://github.com/makeumitchel/measuring-the-degree-of-model-based-control'.

Redish, V., Jensen, S. & Johnson, A. (2008), 'A unified framework for addiction: vulnerabilities in the decision process.', *The Behavioral and brain sciences* **31**(4), 415–487.

Rummery, G. & Niranjan, M. (1994), 'On-line q-learning using connectionist systems.'.

Shahar, N., Moran, R., Hauser, T. U., Kievit, R., M. D. & Moutoussis, M. (2019), 'Credit assignment to state-independent task representations and its relationship with model-based decision making.', *Proc Natl Acad Sci U S A* **16**(32), 15871–15876.

Toyama, A., Katahira, K. & Ohira, H. (2019), 'Reinforcement learning with parsimonious computation and a forgetting process.', *Front. Hum. Neurosci.* .

Voon, V., Derbyshire, K. & Rück, C. e. a. (2015), 'Disorders of compulsivity: a common bias towards learning habits.', *Mol Psychiatry* **20**, 345–352.

Vulkan, N. (2000), 'An economist's perspective on probability matching.', *Journal of Economic Surveys* **14**, 101–118.

5 Appendix

Algorithm 1 exhaustive MB

procedure EXHAUSTIVE MB(parameters, data)
  $[\beta_{low}, \beta_{high0}, \beta_{high1}, lr]$ = parameters
  Retrieve $Tm0$, $Tm1$ from data ▷ Transition matrices
  Initialize
  $Q2_{MF} \leftarrow$ nul matrix dim=(3,1)
  LL_{low} , LL_{high0}, $LL_{high1} \leftarrow 0$
  for each trial **do**
    Retrieve [$state0$, $stims1$, $choice0$, $state1$, $stims1$, $choice1$, $state2$, r] from trial
    if high effort **then**
      $Q1_{MB} \leftarrow Tm1(stims1_{high})$*$Q2_{MF}$
      $Q0_{MB} \leftarrow Tm0(stims0)$*max($Q1_{MB}$)
      $action \leftarrow$ find index ($choice0$=$stims0$)
      $LL_{high0} \leftarrow LL_{high0} + \beta_{high0} * Q0_{MB}(action)$ - $log(\sum \exp \beta_{high0} * Q0_{MB})$
      $action \leftarrow$ find index ($choice1$=$stims1$)
      $LL_{high1} \leftarrow LL_{high1} + \beta_{high1} * Q1_{MB}(action)$ - $log(\sum \exp \beta_{high1} * Q1_{MB})$
    end if
    if low effort **then**
      $Q1_{MB} \leftarrow Tm1(stims1)$*$Q2_{MF}$
      $action \leftarrow$ find index ($choice1$=$stims1$)
      $LL_{low} \leftarrow LL_{low} + \beta_{low} * Q1_{MB}(action)$ - $log(\sum \exp \beta_{low} * Q1_{MB})$
    end if
    Update
    $dQ2 \leftarrow r$ - $Q2_{MF}(state2)$
    $Q2_{MF} \leftarrow Q2_{MF} + lr$*$dQ2$
  end for
  if fitting procedure **then**
    $LL \leftarrow LL_{low} + LL_{high0} + LL_{high1}$
    Return LL
  else ▷ Normalized likelihoods
    LL_{low}, LL_{high0}, $LL_{high1} \leftarrow LL_{low}/n_{low}$, LL_{high0}/n_{high}, LL_{high1}/n_{high}
    Return LL_{low}, LL_{high0}, LL_{high1}
  end if
end procedure

Algorithm 2 exhaustive MBMF for data production

procedure EXHAUSTIVE MBMF(parameters, rewards, effort, task)
  Initialize $Q0_{MF}$, $Q1_{MF}$, $Q2_{MF}$
  Retrieve $Tm0$, $Tm1$ from task ▷ Transition matrices
  Retrieve $stims1_{low}$, $stims1_{high}$ from task ▷ Stimuli of the low and high middle stages
  Initialize *output* as an empty cell
  $[\beta_{low}, \beta_{high0}, \beta_{high1}, lr, \lambda, w_{low}, w_{high0}, w_{high1}]$ = parameters
  for each trial **do**
    if high effort **then**
      $stims0 \leftarrow$ random [x1,x2]$\in \{1,2,3\}$, x1$\neq$x2 ▷ 2 spacestations available
      $Q1_{MB} \leftarrow Tm1(stims1_{high})$*$Q2_{MF}$
      $Q0_{MB} \leftarrow Tm0(stims0)$*max$(Q1_{MB})$
      $Q0 \leftarrow w_{high0}$*$Q0_{MB}$+(1-w_{high0})*$Q0_{MF}(stims0)$
      $Ps \leftarrow$ Softmax$(\beta_{high0}$*$Q0)$
      $action \leftarrow$ random x$\in \{1,2\}$ following Ps
      $choice0 \leftarrow stims0$(action) ▷ pick 1 spacestation
      $state1 \leftarrow$ index of $(Tm0(choice0)=1)$
      $stims1 \leftarrow stims1_{high}(state1)$ ▷ 2 spaceships available
      $Q1 \leftarrow w_{high1}$*$Q1_{MB}(stims1)$+(1 - w_{high1})*$Q1_{MF}(stims1)$
      $Ps \leftarrow$ Softmax$(\beta_{high1}$*$Q1)$
      $action \leftarrow$ random x$\in \{1,2\}$ following Ps
      $choice1 \leftarrow stims1$(action) ▷ pick 1 spaceship
    end if
    if low effort **then**
      $state1 \leftarrow$ random x$\in \{1,2\}$
      $stims1 \leftarrow stims1_{low}(state1)$ ▷ 3 spaceships available
      $Q1_{MB} \leftarrow Tm1(stims1)$*$Q2_{MF}$
      $Q1 \leftarrow w_{low}$*$Q1_{MB}$+(1 - w_{low})*$Q1_{MF}(stims1)$
      $Ps \leftarrow$ Softmax$(\beta_{low}$*$Q1)$
      $action \leftarrow$ random x$\in \{1,2,3\}$ following Ps
      $choice1 \leftarrow stims1$(action) ▷ pick 1 spaceship
    end if
    $state2 \leftarrow$ index of $(Tm1(choice1)=1)$ ▷ pick 1 alien
    $r \leftarrow rewards(trial, state2)$
    Update
    if high effort **then**
      $dQ0 \leftarrow Q1_{MF}(choice1)$ - $Q0_{MF}(choice0)$
    end if
    $dQ1 \leftarrow Q2_{MF}(state2)$ - $Q1_{MF}(choice1)$
    $dQ2 \leftarrow r$ - $Q2_{MF}(state2)$
    if high effort **then**
      $Q0_{MF} \leftarrow Q0_{MF} + lr$*$dQ0 + \lambda$*$lr$*$dQ1 + \lambda^2$*$lr$*$dQ2$
    end if
    $Q1_{MF} \leftarrow Q1_{MF} + lr$*$dQ1 + \lambda$*$lr$*$dQ2$
    $Q2_{MF} \leftarrow Q2_{MF} + lr$*$dQ2$
    Store responses
    $output \leftarrow [state0, stims1, choice0, state1, stims1, choice1, state2, r]$
  end for
  Return *output*
end procedure