Better Than Maximum Likelihood Estimation of Model-Based And Model-Free Learning Style

Sadjad Yazdani (sajjad.yazdani@gmail.com)
University of Tehran
Abdol-Hossein Vahabie
University of Tehran
Babak Nadjar Araabi
University of Tehran
Majid Nili Ahmadabadi
University of Tehran

Research Article

Keywords: Model-Based and Model-Free Combined Learning, modeling different styles of learning, k Nearest Neighbors estimation versus model fitting, Methods of analysis behavioral observation, Behavioral Parameter extraction method

Posted Date: September 9th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-880233/v1

Better than maximum likelihood estimation of model-based and model-free learning style

Sadjad Yazdani[1*], Abdol-Hossein Vahabie[1,2,3], Babak Nadjar Araabi[1], Majid Nili Ahmadabadi[1]

1- Cognitive Systems Laboratory, Control and Intelligent Processing Center of Excellence (CIPCE), School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran

2- Department of Psychology, Faculty of Psychology and Education, University of Tehran, Tehran, Iran

3- School of Cognitive Sciences, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran

Email: sajjad.yazdani@ut.ac.ir

Address: Room 403, Machine Learning and Computational Modeling (MLCM) Lab, School of Electrical and Computer Engineering, College of Engineering, University of Tehran, North Kargar St., Tehran, Iran.

Phone number: +98-913 293 2397

Abstract:

Various decision-making systems work together to shape human behavior. Habitual and goal-directed systems are the two most important ones that are studied by reinforcement learning (RL), using model-free and model-based learning methods, respectively. Human behavior resembles the weighted combination of these two systems. Such a combination is modeled by the weighted sum of action values of the model-based and model-free systems. The weighting parameter has been mostly extracted by "maximum likelihood" or "maximum a-posteriori" estimation methods. In this study, we show these two well-known methods bring many challenges, and their respective extracted values are less reliable, especially in the case of limited sample size or at the proximity of extremes values. We propose that using *k*-nearest neighbor, as a free format estimate, can improve the estimation error. *k*-nn uses global information extracted from the behavior such as stay probability, along with fitted values. The proposed method is examined by simulated experiments, where obtained results indicate the advantage of our method in reducing both bias and variance of the error. Investigation of the human behavior data from previous studies shows that the proposed method results in more statistically robust estimates in predicting other behavioral indices such as the number of gaze directions toward each target or symptoms of some psychiatric disorders. In brief, the proposed method increases the reliability of the estimated parameters and enhances the applicability of reinforcement learning paradigms in clinical trials.

Keywords: Model-Based and Model-Free Combined Learning; modeling different styles of learning; k Nearest Neighbors estimation versus model fitting; Methods of analysis behavioral observation; Behavioral Parameter extraction method

1 Introduction

It is believed that multiple cognitive systems control human decision-making. Habitual and goal-directed systems are responsible for most decisions and learnings during the human's lifetime [1,2]. The habitual system constructs habits and automatic decisions, and the goal-directed system is mostly involved in the planning. Reinforcement learning researchers ascribe the habitual and goal-directed systems to Model-Based (MB) and Model-Free (MF) learning styles, respectively. In MB learning, an environmental model is recruited to evaluate each choice in the current state; on the other hand, in the MF style, the choice-values update without considering any explicit model of the environment and the value of each action at each state learn by trial and error. In both styles, the estimated value of actions affects decision making, and only the way action values are determined is different.

Previous studies have shown that individuals use a combination of MB and MF learnings to guide their behavior during learning tasks [2–7]. It has been proved that the hybrid model is a good descriptor of subjects' behavior, in which subject model run both MB and MF algorithms in parallel and make choices according to a weighted combination of the action values. Through this model, which is clarified in the model structure section, the combination weight (w) is the parameter that determines the subject's preference towards MB and MF styles [8]. To simplify, researchers usually assumed this parameter constant for each subject throughout the task but can vary across subjects[8–10].

Computational models help to isolate distinct cognitive components underlying maladaptive behavior, while the model parameters associated with those components can be used to understand the potential cognitive deficit sources [11]. The parameter that determines the subject's preference towards MB and MF styles, is among such elements that can be used to assess and diagnose psychiatric disorders and evaluate the effectiveness of treatments [12].

Traditionally, the fitting methods (maximum likelihood, maximum a posteriori) are used to find out the subject's preference towards MB and MF styles from observation of his/her choices. The most important algorithm in model fitting is Maximum Likelihood. In case that there is no information except behavioral observation the likelihood is the best objective function for model fitting, this method is based on the idea that the observed data is more likely to have come about as a result of a particular set of parameters. Also, the "Maximum A Posteriori" (MAP) method is used in case we have any prior knowledge of parameters. In this study, we want to use other information, especially global information that can be simply caught from observation, in the process to find out the weighted combination of the action values. We propose that using a data-driven learning method besides traditional fitting methods can help to improve the estimation accuracy and reliability. In this study, we use the k-nearest neighbor algorithm as a simple learning method. Other learning algorithms like deep neural network also can be recruited for the same purpose.

Consideration of changes in the subject's preference towards MB and MF styles (w) due to pharmacological or cognitive manipulations or neuropsychiatric conditions will provide important insights for clinical research. For example, Over-reliance on MB style could lead to inflexible decisions in addiction and compulsion [13–15]. Some studies show that patients with obsessive-compulsive disorder (OCD) prefer the MF learning style more than MB one [16–19]. Wit et al. show that mild Parkinson's disease has led to impaired MF habit formation [20]. Also, Culberth et al. show that in schizophrenic patients MB behavior is reduced [21].

On a broader view, there is a growing consensus that computational modeling can be constructive to understand psychiatric disorders. Therefore, reliable and precise estimation of the combination weight (w) is important for many applications. However, reliable estimation

of parameters is a challenge due to the noise in behavior and confounding factors as well as low sample size, especially for extreme values, which are more likely in pathologic conditions.

In many cases, the model fitting results in extreme values for combination weight due to the limited sample size and biases towards boundaries [19]. The analysis shows that the precision of the estimated subject's preference towards MB and MF styles by traditional model fitting is not excellent, and it is biased toward pure MF, especially when the other parameters of the model are not in the appropriate range. Simulations show that the low value of learning rate or high value of the Boltzmann machinery temperature results in more significant error in the estimation of combination weight in model fitting.

In the following, Section 2 reviews the basic model architecture (Section 2.1) followed by the proposed method that makes clear the difference between the proposed method and the traditional fitting method (Section 2.2 and 2.3) also reviews the k-nn estimator (Section 2.4). Section 2.5, 2.6, and 2.7 demonstrate details about implementation. In Section 3 We first test the performance of the traditional fitting method in the variation of some model parameters (Section 3.1 and 3.2) followed by setting parameters of the proposed method (Section 3.3), then we analyze the result of the proposed method (Section 3.4). At the end of Section 3, we analyze the effect of noise in extracting the w (Section 3.5). Section 4 demonstrates the performance of the k-nn method in experimental data and the advantage of the proposed method. The last section of this paper discusses the proposed method and sum up the results in the conclusion (Section 5).

2 Method

2.1 The Daw task

Daw et al. for dissociating the MB and MF components of human behaviors develop a two-step task [8]. Numerous other researchers have used this or a variation of this task[10,22,31,23–30]. Fig 1 illustrates the Markov Decision Process (MDP) model for this task.

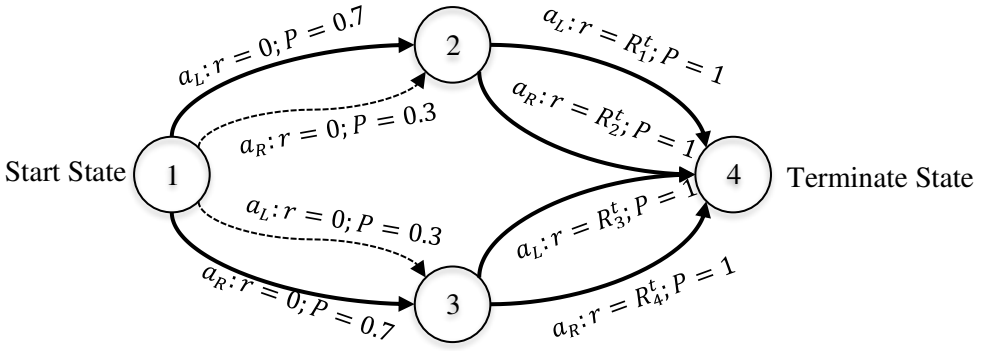

Fig 1. Daw Task MDP model: In all non-terminate states, two different actions (labeled as a_L and a_R) are available. In the first state, each action is predominantly associated (with a 70% probability) with one of the second level states. The transitions with 70% probability forenamed **Common**, and those with a 30% probability named **Rare**. Any action in second-level states are associated with different reward probabilities that fluctuate independently across the session by a random walk (with the standard deviation of step size: 0.1) limited between 0.25 and 0.75. In any trial, the subject has to decide about two actions. The first action has no reward, and the second one results in the rewarded or unrewarded trial. Thus, subjects have to make trial-by-trial adjustments in their choice to maximize the probability of achieved reward.

2.2 Model Structure

Based on [8], subjects run both MB and MF learning style in parallel and make choices according to the linear weighted combination of the action values that come from MB and MF systems. This hybrid model has also been used by several other researchers[30,32,33]. Fig 2 shows the flowchart of this model, the parameters of the model, and available observations from the human task for each section. In this model, the decisions are made probabilistically based on

the values assigned to available actions in the specific state (Q (S, a)), and these values are updated at each trial.

In any trial (t), the value of each action (a) of the first state calculates by the weighted sum of MB (Q^{t}_{MB}) and MF (Q^{t}_{MF}) system value (weight: w) according to the equation (1).

$$Q^{\mathrm{t}}(1,\mathrm{a}) = w \times Q^{\mathrm{t}}_{MB}(1,\mathrm{a}) + (1-w) \times Q^{t}_{MF}(1,a) \quad (1)$$

The *stickiness* increases the value of the previous action for the current trial by adding P to its value (equation (2)).

$$\hat{Q}^{t}(1,a) = \begin{cases} Q^{t}(1,a) + P & if\ a\ is\ the\ previous\ action \\ Q^{t}(1,a) & otherwise \end{cases} \quad (2)$$

The Boltzmann machine is a stochastic, biologically-plausible approximation of the maximum operation [34], which is widely used to extract the probability of choosing each action based on their values(equation (3)).

$$P\left(\mathrm{a};\hat{Q}^{\mathrm{t}}\right) = \frac{e^{\beta \times \hat{Q}^{t}(1,a)}}{\sum_{\hat{a}} e^{\beta \times \hat{Q}^{t}(1,\hat{a})}} \quad (3)$$

β is the inverse temperature that controls the trade-off between exploitation and exploration. Due to the non-deterministic environment and its probabilistic nature for rewards, it is usually assumed as a fixed parameter over trials but differs across subjects[8].

In the second stage of the task, corresponding Q values in each state determine the probability of chosen action by the same stickiness and Boltzmann machinery.

In the beginning, the Q values of all state-actions initialized to zero, and the update rules (equation (4)) changes the value of state-actions at the end of each trial.

In the second stage of the task, the update rule is the same for both MB and MF approaches. In the first stage, however, action values are updated by State–action–reward–state–action

(SARSA)-λ for the MF method, while the model of the environment is used to update the Q_{MB}. Note that update rules for Q_{MF} and Q are applied only to the performed action, while Q_{MB} updates all action values of the first stage.

$$Q_{MF}^{t+1}(1,a) = Q_{MF}^{t}(1,a) + \alpha_1\left(Q_{MF}^{t}(1,a) - Q^{t}(S,a)\right) + \lambda\alpha_1\left(r^{t} - Q^{t}(S,a)\right)$$

$$Q^{t+1}(S,a) = Q^{t}(S,a) + \alpha_2(r^{t} - Q^{t}(S,a))$$

$$Q_{MB}(1,a) = \sum_{S}^{2,3} \mathrm{P}_T(1,a,S) \times \max_{\hat{a}} Q^{t}(S,\hat{a}) \quad (4)$$

where $P_T(1,a,S)$ is the probability of transition from state 1 towards the second step state S acting a and may be assumed as the real value (i.e., 0.3 and 0.7) or calculated by Beta-Binomial Bayesian updating rule according to equation (5) [9].

$$P_T(1,\mathrm{a},S) = \frac{1+N(1,a,S)}{2+\sum_{\hat{S}} N(1,\mathrm{a},\hat{S})} \quad (5)$$

where $N(1, a, S)$ is the number of transitions from start-state to state S by acting the a. In this study, we calculate the probability of transition according to equation (5).

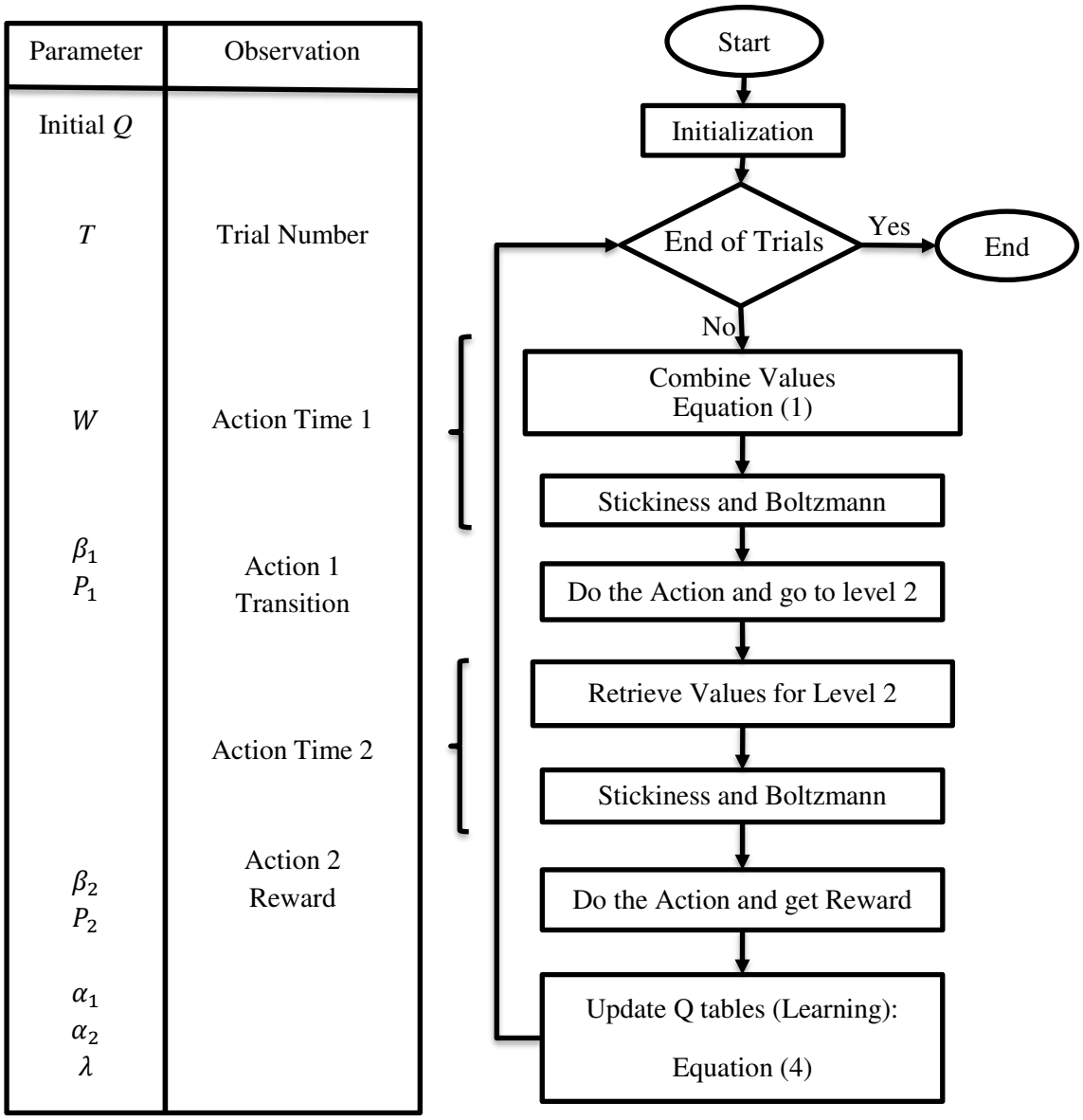

Fig 2. The hybrid model for reinforcement learning: The flowchart presents the model for the reinforcement learning process by combine MB and MF styles. The parameter box specifies the parameters used in each part of the model. Also, the observation box specifies the available observation from the behavior for each part of the model.

The *Q*s values are set to zero initially, and the stop criterion is the fixed number of trials, *T*, which is set to 201 similar to [8]. We introduced the model in its most general form above; however, some studies like [19,35] use the reparameterization method and alternative models with different free parameters. Also, some studies like [36] use the response time in a model which is

not available for our simulation. By setting some parameters to a fixed value or identical in two stages, nine versions of the model extracted. Table 1 lists these models and clarifies subsets of parameters for each version. These models are nested, and the 8Param version is the most complicated one.

Table 1. Comparison of model versions: nine versions of the general model introduced by fixing some parameters to a fixed value or identical in two stages.

Parameter Name	Combination weight of MB/MF	Learning Rate 1st Step	Learning Rate 2nd Step	Inverse Temperature 1st Step	Inverse Temperature 2nd Step	Eligibility Trace	Stickiness to repeat the same 1st action	Stickiness to repeat the same 2nd action
Parameter Symbol / Version	w	α_1	α_2	β_1	β_2	λ	P_1	P_2
3ParamV1	w	α	α	β	β	1	0	0
3ParamV2	w	α	α	β	β	0	0	0
4Param	w	α	α	β	β	λ	0	0
5ParamV1	w	α_1	α_2	β_1	β_2	1	0	0
5ParamV2	w	α_1	α_2	β_1	β_2	0	0	0
6Param	w	α_1	α_2	β_1	β_2	λ	0	0
7ParamV1	w	α_1	α_2	β_1	β_2	λ	P	P
7ParamV2	w	α_1	α_2	β_1	β_2	λ	P	0
8Param	w	α_1	α_2	β_1	β_2	λ	P_1	P_2

2.3 The *w* estimating

Model fitting extracts the parameter that determines the subject's preference towards MB and MF styles (w) by maximizing the similarity between the model decision and human behavior, so both the model and objective function affect the performance of model fitting.

Theoretically, the likelihood is the best objective function for model fitting, in case that there is no information except behavioral observation. This method is based on the idea that the observed data is more likely to have come about as a result of a particular set of parameters. This method is used widely in behavioral sciences[37]. Also, the "Maximum A Posteriori" (MAP) method used in case we have any prior knowledge of parameters. In this study, the objective functions are minimized by the interior-point optimization algorithm, and five different random start points are used to have more chance of global optimization.

In this paper, we use other available information, including behavior statistics and indices, besides the fitted values of the parameter to extract the *w* more precisely. Although this study focuses on the action selection observation, the estimator can be more precise by using some other measurable parameters like confidence level or response time [36]. In the proposed method, we use a *k*-nn estimator (also known as k-nn regressor) as a learning system to extract the *w* from behavior. *k*-nn is a supervised, free format learning method, and has been widely used as a good point estimator [38,39]. *k*-nn uses a dataset of labeled feature vectors to estimate the *w* parameter. We generate the dataset by simulation of the model, which has been assumed to be used by subjects.

Fig 3 illustrates the overall structure of both model fitting and the learning algorithm to extract the *w* parameter from observed behavior. The model fitting needs model and objective function, and as mentioned before, both of them have effects on performance.

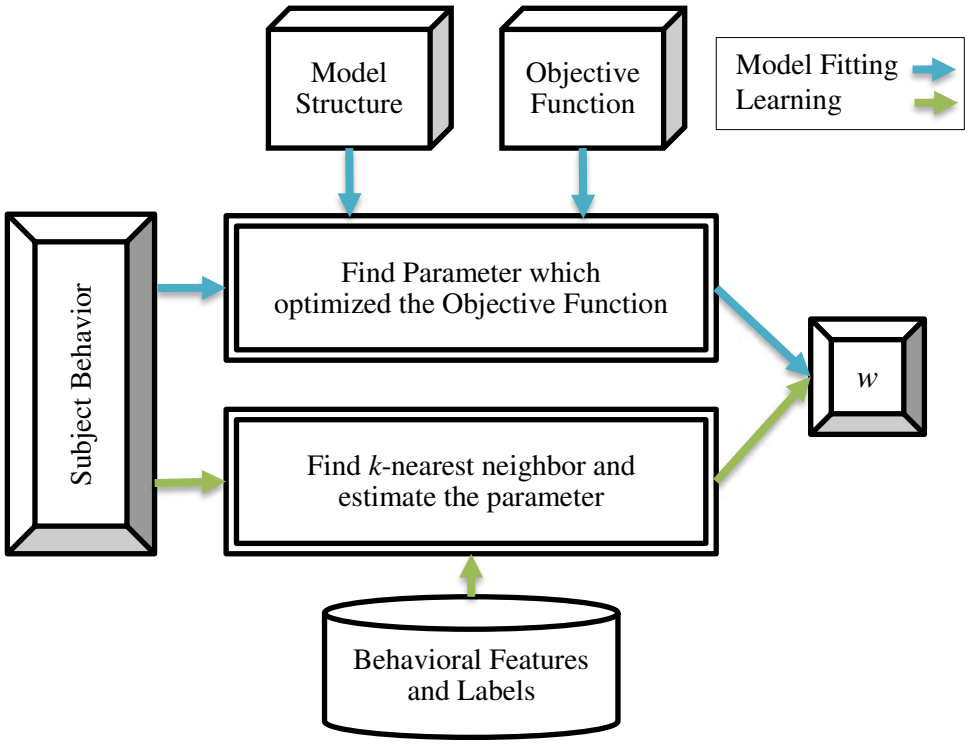

Fig 3. Method for extracting the combination weight (w): Both model fitting and learning methods get the observed behavior of the subject, and the main output is the w (the weight of the combination of the MB and MF learning style in the Daw task). In the model fitting method, both the assumed model and objective function have effects on performance. The learning method needs a database of features and labels, which is more robust to noise and errors rather than model structure and objective function.

The k-nn method extracts the w based on a dataset that is generated by simulation. We use Euclidean distance in normalized feature space to find k-nearest neighbors. The estimation value is the sum of k-nearest neighbor values weighted by the inverse of Euclidean distance. Note that due to partial observation of the action values, there is always an error in the extraction of w, even if we know the exact behavioral model.

2.4 k nearest neighbor estimation

As mentioned before, the k-nn is a supervised learning method. Fix and Hodges wrote a technical report in 1951, including a method for pattern classification that has since become known as the k-nearest neighbor rule [40]. Later in 1967, Cover and Hart have shown that this

classification method's error is bounded above by twice the Bayes error rate (T. M. Cover & Hart, 1967). One year later, Cover extend this bounding to the estimation method[41]. Such formal properties of *k*-nn start a long line of investigation, including distance weighted approaches[42], which we used in this study.

Assume that we have some exact value of the combination-weight of MB/MF learning styles (*w*), for numerous observations. The features vector (See next section for details) extracted from each observation and these feature vectors and related *w* values (well-known as the label) are stores in a database. By having this database, we want to estimate the *w* of newly observed behavior data. Based on this data, the feature vector was calculated, then we linearly normalized this vector, based on feature space. Then we calculate the Euclidean distance (*d*) of this feature vector from other vectors in the features space and *k* nearest features found based on *d*. now the estimated value of combination weight for this observation is given by equation (6)

$$\widehat{w_0} = \sum_{i \in N_0} v_i \times w_i \,/\, \sum_{i \in N_0} v_i \tag{6}$$

Where w_i is the label of *i*th sample in dataset and N_0 is the index set of *k* nearest neighbor, also the weighted factor v_i is calculated by equation (7).

$$v_i = \begin{cases} \frac{d_{max} - d_i}{d_{max} - d_{min}} & if\ d_{max} \neq d_{min} \\ 1 & Otherwise \end{cases} \tag{7}$$

In equation (7) the d_{max} and d_{min} are the maximum and minimum distance values of neighbor respectively.

The feature space and hyperparameter *k* have the most impact on the results. The *k* parameter controls the localization and generalization of the *k*-nn learning method and has an optimal value. For the proposed method we found the optimized *k* by exhaustive search. To

optimize the feature space, we used forward selection. Forward selection begins with an empty feature set and in each step, the one feature that gives the best improvement is added to feature space. These steps continue till all features were added or adding other features makes no improvement.

2.5 Features

The "stay probabilities" is the first group of features that we use in *k*-nn. The stay probabilities calculate by counting the stays, i.e., choosing the same action as the previous trial in the first stage, from the observed behavior. This feature calculates in different conditions that are related to the reward value (either Rewarded or Unrewarded) and transition (either Common or Rare) of previous trials. Numerous studies on the same Daw task use this measure[8,43]. Also, the selection of the "Best or Not Best" decision in the first stage is another condition for stay probability. The best decision is the one that the common transition changes state toward the most probable reward which is defined from the task setting. We use the stay probability in all situations across three different conditions, as listed in Table 2 from 1 to 27.

Furthermore, the slope of stay probabilities, as indices for MF (equation (8)) and MB (equation (9)) behavior [44], were also used as another behavioral indicator in feature space.

$$\mathrm{I}_{MF}^{\mathrm{PStay}} = P\left(S \,\middle|\, Re, C\right) + P(S|Re, R) - P(S|Ur, C) - P(S|Ur, R) \tag{8}$$

$$\mathrm{I}_{MB}^{\mathrm{PStay}} = P\left(S \,\middle|\, Re, C\right) - P(S|Re, R) - P(S|Ur, C) + P(S|Ur, R) \tag{9}$$

Table 2. Features extracted statistically based on stay probability

#	Symbol	Description
1	*P(S)*	Stay Probability over all trials
2	*P(S|Re)*	Stay Probability over trials after the Rewarded trial

3	$P(S\|Ur)$	Stay Probability over trials after the Unrewarded trial
4	$P(S\|C)$	Stay Probability over trials after the Common trial
5	$P(S\|R)$	Stay Probability over trials after the Rare trial
6	$P(S\|B)$	Stay Probability over trials with the Best decision
7	$P(S\|NB)$	Stay Probability over trials with Not Best decision
8	$P(S\|Re,C)$	Stay Probability over trials after different situations across Rewarded, Unrewarded, Common and Rare of the previous trial as well as Best or Not Best decision.
9	$P(S\|Re,R)$	
10	$P(S\|Ur,C)$	
11	$P(S\|Ur,R)$	
12	$P(S\|B,C)$	
13	$P(S\|B,R)$	
14	$P(S\|NB,C)$	
15	$P(S\|NB,R)$	
16	$P(S\|B,Re)$	
17	$P(S\|B,Ur)$	
18	$P(S\|NB,Re)$	
19	$P(S\|NB,Ur)$	
20	$P(S\|B,C,Re)$	
21	$P(S\|B,C,Ur)$	
22	$P(S\|B,R,Re)$	
23	$P(S\|B,R,Ur)$	
24	$P(S\|NB,C,Re)$	
25	$P(S\|NB,C,Ur)$	

26	$P(S\|NB,R,Re)$	
27	$P(S\|NB,R,Ur)$	
28	I_{MF}^{PStay}	the slope of stay probabilities
29	I_{MB}^{PStay}	

We also use some features that need the model fitting procedure. We added model fitting analysis indices, equation (10) and (11), which is introduced by Miller et al. to the feature space [44].

$$\mathrm{I}_{MF}^{\mathrm{Fit}} = \left(1 - w^{\mathrm{Fit}}\right) \times \beta_1^{\mathrm{Fit}} \tag{10}$$

$$\mathrm{I}_{MB}^{\mathrm{Fit}} = w^{\mathrm{Fit}} \times \beta_1^{\mathrm{Fit}} \tag{11}$$

In these equations, w^{Fit} and β_1^{Fit} are the combination weight and inverse temperature of the first stage respectively, and extracted by best fits according to the Akaike Information Criterion (AIC) by fitting methods, which can be either MLE or MAP.

Feature space also includes estimated MLE and MAP fitting values of some parameters. In sum, ten features extracted by the model fitting procedure were added (Table 3). All these features are extracted by best fits according to the Akaike Information Criterion (AIC).

Table 3. Features by model fitting

#	Symbol	Description
30	I_{MF}^{MLE}	$\mathrm{I}_{MF}^{\mathrm{Fit}} = \left(1 - w^{\mathrm{Fit}}\right) \times \beta_1^{\mathrm{Fit}}$ $\mathrm{I}_{MB}^{\mathrm{Fit}} = w^{\mathrm{Fit}} \times \beta_1^{\mathrm{Fit}}$
31	I_{MB}^{MLE}	
32	I_{MF}^{MAP}	

33	I_{MB}^{MAP}	
34	w^{MLE}	Parameters Extracted by Model Fitting
35	α_1^{MLE}	
36	β_1^{MLE}	
37	w^{MAP}	
38	α_1^{MAP}	
39	β_1^{MAP}	

2.6 Generated Dataset for *k*-nn

As a supervised learning method, *k*-nn needs a training dataset with appropriate labels to function correctly. So, we simulate 80,000 RL independent agents with random parameters and the Daw8Param version (see Table 1), then record their behavioral observations. In this study, all random parameters were sampled according to Table 4 as well as the prior for MAP. Each simulation generates a sequence of trials and related observations, which are all labeled by the parameter that indicates the subject's preference towards MB and MF styles (w). Moreover, the 10-fold cross-validation is used for training in hyper-parameter tunings, i.e., k and feature selection. To remove the estimator bias in extremes, 10000 fully MB agents, and 10000 fully MF agents, are added to the training dataset.

Table 4. Parameters, range, and random values for independent agents. (Beta(.) is the beta distribution)

Parameter Symbol	**Description**	**Low**	**High**	**Random Value**
w	MB/MF combination weight	0	1	uniform
α	Learning Rate	0	1	Beta(1.2,1.2)

β	Inverse Temperature of Soft Max machine	1	10	1+9×Beta(1.2,1.2)
λ	Eligibility Trace	0	1	Beta(1.2,1.2)
P	Stickiness to repeat the action	0	0.2	0.2×uniform

2.7 Model the noise in decision making

It has been shown that the inclusion of the lapse rate for human subjects can improve the fitting quality in many psychophysical paradigms [45]. This lapse rate is due to the trials in which the subject has not attended and responded randomly. We include this possibility for agents in simulations. To do so, each decision of the agent is reversed by a probability called lapse rate or noise level. We simulate the noisy model for different noise levels in [0 0.5] intervals.

3 Results

To have a clear insight, we first analyze the performance of the MLE and MAP method of fitting. We simulate a wide range of human behavior during the Daw task By the Daw8Param model and random parameters. We fit all versions of the model to observation, and then use AIC to choose the best-fitted model version by each fitting method. The error of extracting the combination weight (w) is calculable for these simulations because both values of the agent's w and extracted one are known. We use Mean Absolute Error (MAE) as a point value of the error and standard deviation (STD) to illustrate the distribution of error. The number of agents in the simulations is large enough to have repetitive and trustable results.

3.1 Effect of Agent Parameter Set in Model Fitting

Variation of models in model fitting can lead to different error levels[19], but what about the agent model version? To investigate this effect, we ran 5000 independent agents with different versions listed in Table 1. For each data set, all model versions and both MLE and MAP model fittings are applied to the observation. The result of these fittings by MLE and MAP is summarized in Fig 4-A and B, respectively. Based on this figure, when the agent has zero eligibility trace (either 3ParamV1 or 5ParamV1 that λ=0), but the fitting method assumes a significant eligibility trace (3ParamV2 or 5ParamV2 that λ=1), the estimation error is extensive. Also, the error is substantial in reverse situations. The eligibility trace (λ) controls the effect of the second stage state-action reward on the first stage action value in SARSA-λ machinery. The λ value strongly affects the behavior of pure SARSA-λ. So, by the wrong assumption for λ value, the information about MB-MF in behavior will be confusing and ends in the massive error in the w estimation.

Additional to this remarkable point, knowing the model that the RL agent has used does not always result in a lower error, especially when the model is more complicated. It seems that the randomness of behavior and especially overfitting in the more complex model is the cause of this point.

3.2 Effect of agent learning rate and temperature on model fitting

Assume that an agent uses the 3ParamV1 model while performing the Daw task, and w is extracted by the best model based on the AIC. The question here is whether the parameters' value of the model affects fitting error. To this end, we run 5000 agents with fixed number sets of learning rate (α) and inverse temperature (β) individually while all other parameters are sampled randomly. Fig 5 demonstrates the result of this simulation. Fig 5-A shows that in the

low learning rate, the estimation error is significantly higher. A low learning rate means that the agent cannot follow the changes in the environment, i.e., the changes in the environment are faster than that can be tracked by the low learning rate. This agent has difficulty in choice evaluation by both MF and MB styles. This difficulty ends in wrong decisions that seem random, and model fitting faces more challenges and consequently results in a higher error.

Fig 5-B shows that agents with low Boltzmann inverse temperature (β<3) have high MAE, and it decreases for larger β values. A low value of β means more exploration, and similar to the low value of α results in more like random behavior.

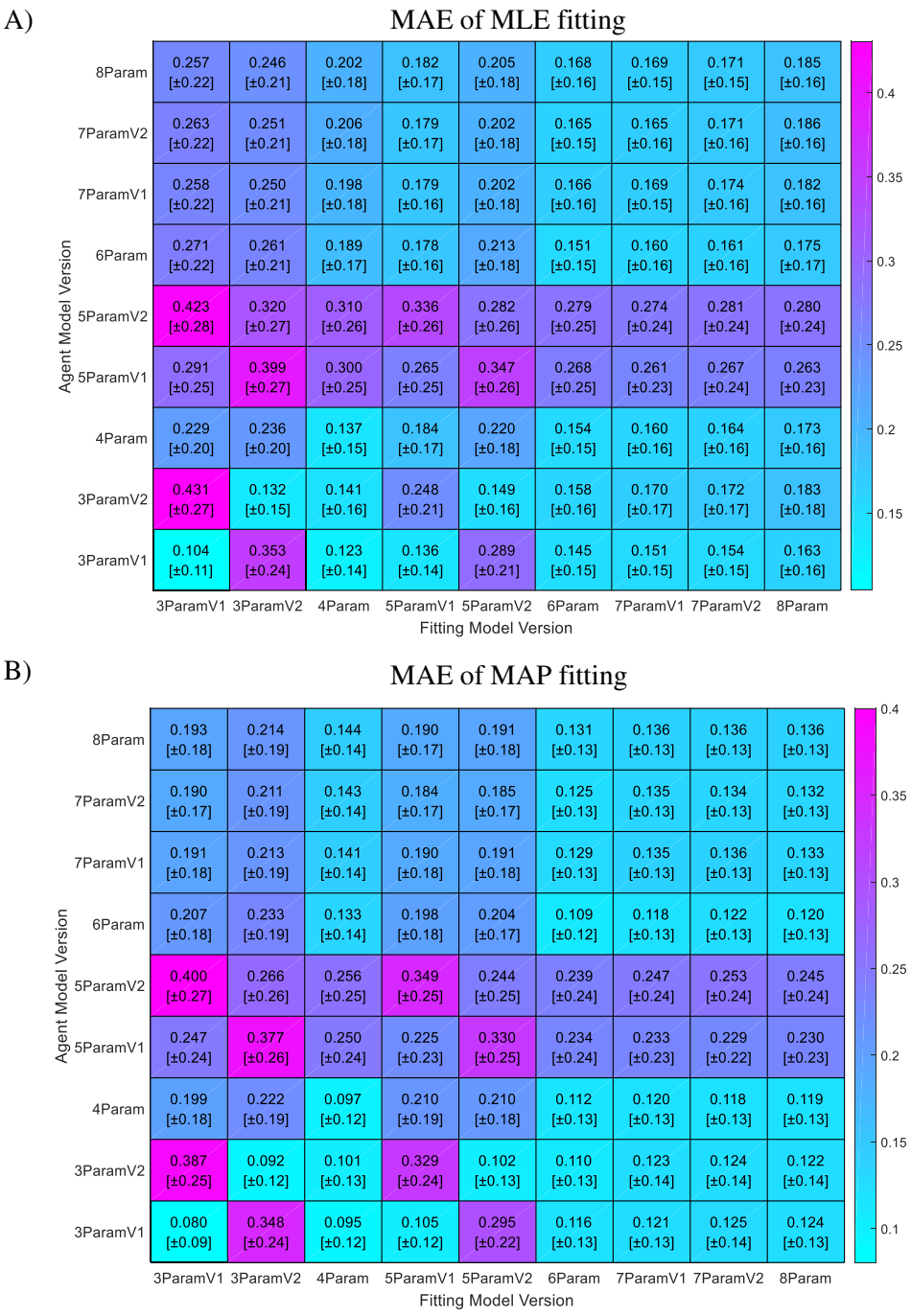

Fig 4. Mean Absolute Error and [STD] of a different model version of fitting by A: MLE and B: MAP versus agent model version. Each row is 5000 agents who performed the task independently, and the column is the result of fitting the model versions to observed behavior.

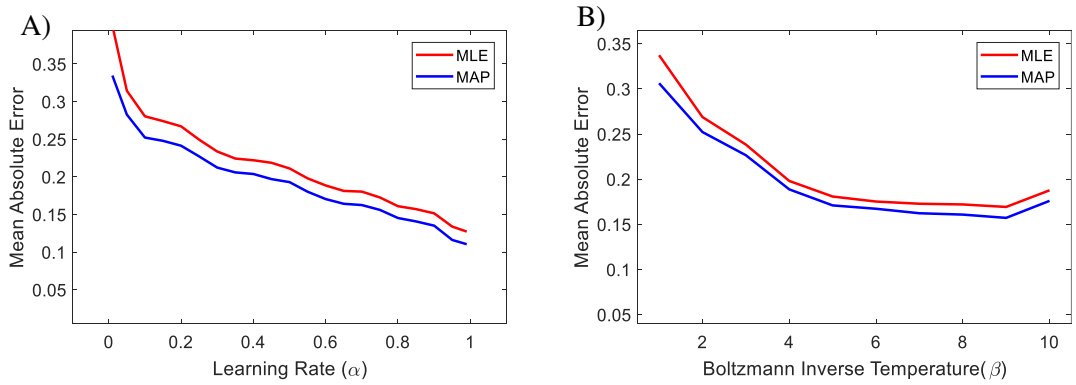

Fig 5. Effect of Agent learning rate (A) and Boltzmann inverse temperature (B) on the performance of fitting: Each point represents 5000 agents that perform the task independently. Agents use the 3ParamV1 model, and other parameters are random. Then both MLE and MAP model fitting used to extract the *w* from observed behavior, and the best model version is selected based on AIC.

3.3 *k*-nn Parameters

The *k* value controls the localization and generalization of *k*-nn. In fact, the low number of neighbors means more localization and more neighbors means more generalization. We know that both localization and generalization have an impact on errors and a trade-off between these two issues is needed to get a good performance. To achieve the best *k*-nn performance, we adapt the *k* value by the exhaustive search to minimize MAE. According to Fig 6.A, when k is greater than 30 and up to 100, the MAE is nearly constant. For these values of *k,* the MAE just varies minimally in the range of 0.1843 to 0.1848, but the optimal value of k is 68, and we use this value in all different situations. In addition to this parameter, the feature selection and feature combination can improve *k*-nn performance.

3.3.1 Feature Selection

Selecting useful features can improve the performance of the *k*-nn estimator. Forward selection is one of the most commonly used feature selection methods. Some features that are introduced need a fitting process and the fitting, facing some challenging issues like

computational load, selection of a good model and optimization algorithm, etc.. To adjust the proposed method for some practical applications in which mentioned factors limit the use of fitted parameters, we can ignore such features and in return reduce the performance (although in some cases like having not a good model or noisy observation this neglecting can improve the performance). We used forward selection in two different situations based on the available information and analytics:

1- All features will be computed (needs model fitting, i.e., MLE and MAP estimation).

2- Features from model fitting are excluded

By these conditions, we achieve two different feature subsets. Based on Fig 6-B, the subset $\wp_{sub1}$= {w^{MAP}, I_{MF}^{PStay}, P(S|Re,C), I_{MF}^{MLE}, I_{MB}^{PStay}, α_1^{MAP}, P(S|Ur), P(S|B,Re), I_{MB}^{MLE}, I_{MF}^{MAP}, P(S|Re), P(S|B,C), P(S|Re,R), P(S|Ur,R), P(S|NB,Ur), P(S|B)} is selected when all features is available. Adding more features increases the MAE for estimation.

Fig 6-C illustrates the forward selection where total fitting-based features exclude from features space. The obtained feature subset is then $\wp_{sub2}$= { P(S|Ur,R), I_{MF}^{PStay}, I_{MB}^{PStay}, P(S|NB,C), P(S|NB,Ur), P(S|NB), P(S|Re), P(S|C), P(S|Ur), P(S|Ur,C), P(S|Re,C), P(S|R), P(S|Re,R), P(S|B,C), P(S|B), P(S|B,C,Ur)}. Adding any feature increases the MAE. When fitting based features are excluded, the computational load and the effect of decision noise in results reduced, although it increases the MAE.

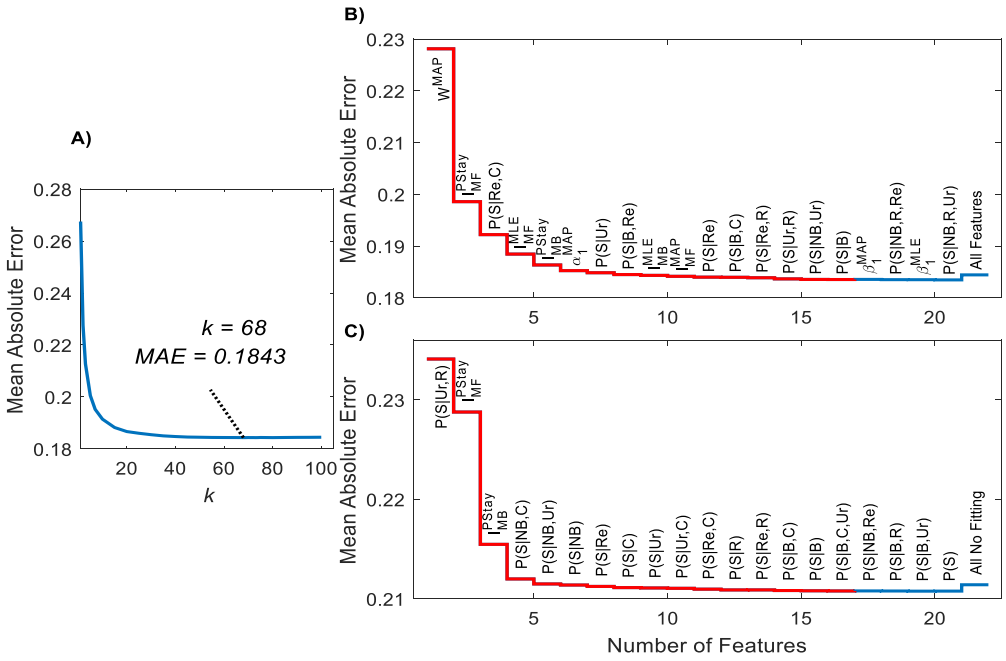

Fig 6. A: Variations of MAE vs. k in k-nn estimator to find the best k. B and C: Forward selection in different conditions. B: All Features included C: Fitting features excluded. The red steps are selected features for which their inclusion in feature space reduces MAE. The 10-fold cross-validation on the k-nn dataset is used.

3.4 Performance

Fig 7 illustrates the scattering of the extracted weight of combination (w) by k-nn estimator and traditional model fitting relative to the corresponding value of agents. To have a good view, we divide it into five different area kinds. As mentioned before, due to partial observation of subjective action value, the exact combination weight (w) cannot be extracted, so a small error (below 0.1) assumes tolerable. Ideally, the tolerable area percentage should be 100. Grouping the subjects based on style learning is an application of extracting the w. So, if an error in w extraction results in the incorrect label for the subject, the error is considerable. Based on Fig 7, the results of k-nn w estimations by using $\wp_{\text{sub1}}$ features are the nearest one to ideal. Also, the MLE and MAP have biased toward the MF style, but k-nn methods solve this problem. Besides, the scatter illustrates that sticking to the extreme values, which is the most problem of traditional fitting methods, is addressed by k-nn.

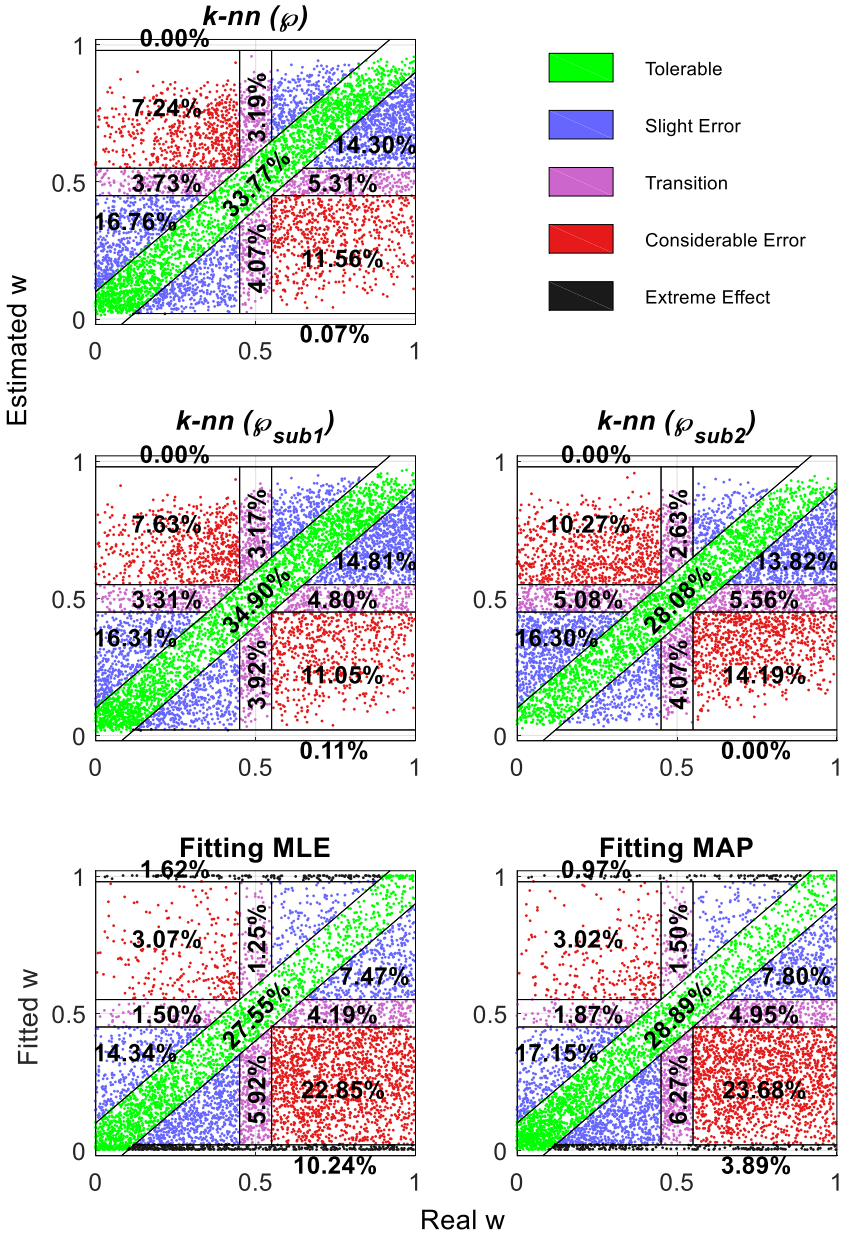

Fig 7. Fitting and Estimation performance. The horizontal axis is the agent's *w,* and the vertical axis is the output of fitting or *k*-nn estimator. The real *w* is the weighting parameter of 10000 independent agents that perform the Daw task by the Daw8param model and random parameters. We fit all versions of the model to observation by both MLE and MAP fitting methods, and then the best version was selected based on the AIC. Also, we apply *k*-nn estimation by different feature spaces, i.e. $\wp$, $\wp_{\text{sub1}}$ and $\wp_{\text{sub2}}$. Each subfigure is divided into different areas. The points that have a small error (below 0.1) are assumed as tolerable errors and scattered in green color. The considerable error area (red points) are those that the dominant strategy changed between MB to MF and blue points that the dominant strategy did not change considered as a slight error. Those areas that did not have a dominant strategy (extracted or agents *w* is not greater than 0.55 nor smaller than 0.45) have been assumed as the transition areas (magenta color). The top and bottom areas are those points that the extracted *w* stick to the extreme and scattered in black color. Distribution of the points, clarified by percentage in any area.

As mentioned before, the individual difference is an important issue, especially in computational psychiatry. In many cases, the percentage of high error is more important than the exact estimation; in other words, it is crucial to have an estimate with low error variance. This issue is addressed by bias-variance tradeoff in the literature. Fig 8 illustrates error distribution by the normalized histogram of error, the difference between estimated and real values.

Based on Fig 8, both bias and variance of error decrease by the *k*-nn method. For the *k*-nn method, the tail of the distribution is shorter and includes lower values at the tail. It means that the variance of the error is small, and the calculated value of standard deviation (STD) confirms this (Table 5). On the other hand, for the *k*-nn methods, the probability of tolerable error (errors between -0.1 and 0.1) is higher than just fitting methods. Also, Table 5 by reporting MAE and correlation coefficient, confirms that the *k*-nn estimation reduces the bias and variance of error.

Extreme errors in both ML and MAP are relatively higher than *k*-nn based methods. Since it is possible to have extreme values for subject's preference towards MB and MF styles in clinical conditions, these regions are more important. *k*-nn methods correct these errors and make the model more robust in clinical trials. The skewness of the error of fitted combination weight (w) in Fig 8 illustrates a bias toward MF, which is in line with Toyama's study [19].

Table 5. MAE, STD, and R of w extraction error by *k*-nn method and model fitting. It is clear that both bias and variance of error decrease by *k*-nn and the linear regression coefficient increase.

Extraction Method	MAE	STD	R
k-nn ($\wp$)	0.1917	0.1490	0. 5966
k-nn ($\wp_{Sub1}$)	0.1899	0.1493	0.6026
k-nn ($\wp_{Sub2}$)	0.2208	0.1566	0.4221
MLE	0.2699	0.2207	0.4509
MAP	0.2547	0.2116	0.4608

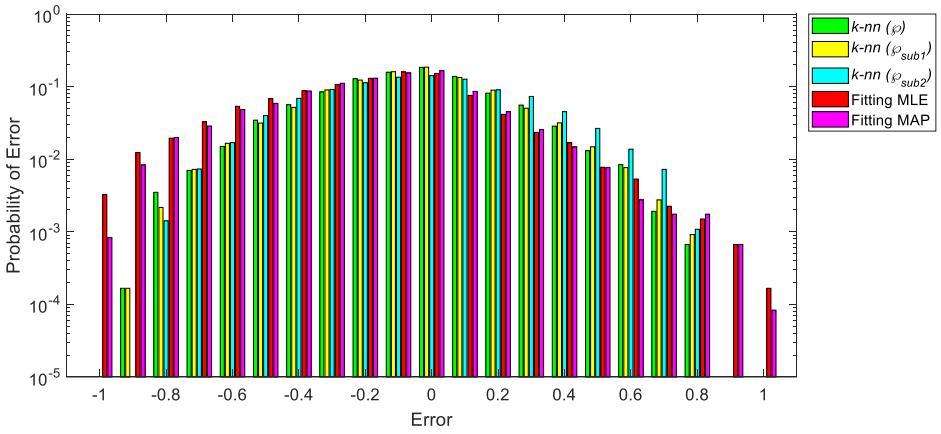

Fig 8. Distribution of error by different models for w extraction. This analysis was done by 10000 independent agents that performed the Daw task by the Daw8param model version and random parameters. After extracting w by all mentions methods, the estimation error (extracted value minus the real one), is calculated. The extracted w is given based on the AIC model comparison of the model version by a different method of fitting also the output of k-nn calculated in different feature spaces of $\wp$, $\wp_{\text{sub1}}$ and $\wp_{\text{sub2}}$.

3.5 lapse in decision making

One practical issue in parameter estimation on human choice data is mistakes in choosing the intended option due to attentional lapse or other issues. These lapses are more critical in clinical applications that due to some disorders lapse rate can be higher than normal people. One aspect that should be considered for the usefulness of an estimation method and its applicability is the robustness of the method in confronting these lapse rates. To investigate the effect of human mistakes, we simulate the model with different lapse rates or noise-ratio. Fig 9 illustrates the difference between traditional fitting methods and the k-nn estimation method.

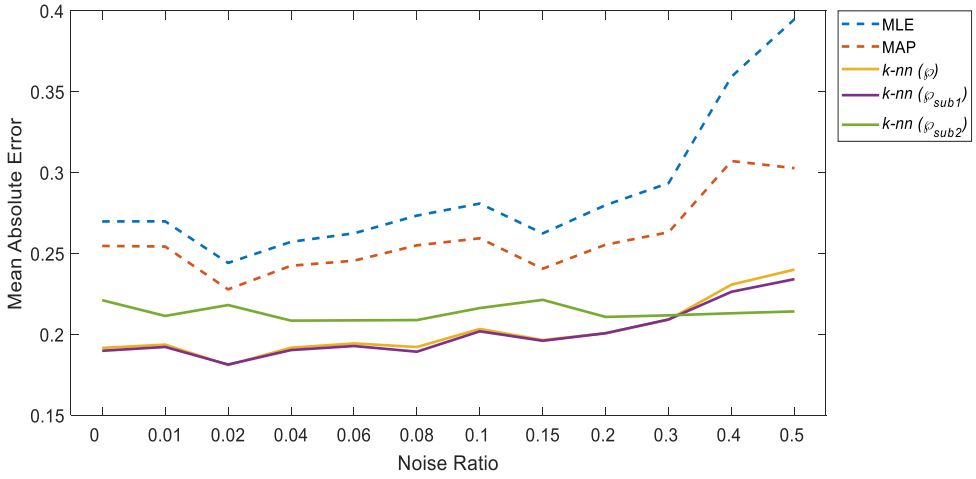

Fig 9. MAE of extracted *w* by *k*-nn and Fitting in the presence of noise. Each point represents 10000 independent agents that perform the task by the Daw8Param and random parameters. After making a decision, it became toggle by the probability of noise ratio. The fitted model is chosen based on AIC. The *k*-nn estimation is applied by all the different feature spaces mentioned before.

Based on Fig 9, it is clear that the *k*-nn methods are more robust than traditional fitting methods against noise or lapse, especially when we exclude the fitting-based features. In fact, the use of statistically extracted features in the *k*-nn method makes it more robust against noise or lapse rate that appear as noise in observation.

4 Experimental data analyses

This section checks the proposed method in real-world experimental data. The data from two different studies have been chosen for validation of the proposed method and the analysis of results based on the combination weight from the proposed method in comparison with the results based on the combination weight from traditional fitting methods shows the superiority of the proposed method. The proposed method reveals some information from the data that could be missing by using traditional fitting methods.

4.1 Analysis of relationship between Learning style and Gaze direction

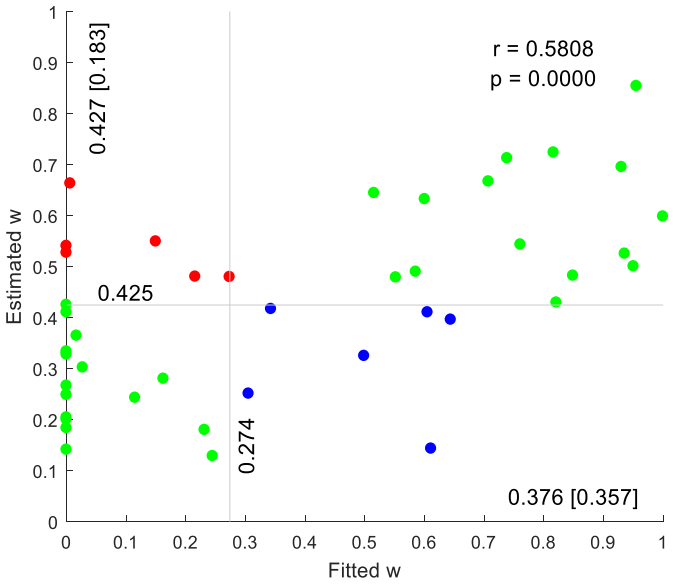

Fig 10 Estimated *w* vs. Fitted *w*. The green points represent those subjects that are in the same group by fitting and estimation. The red spots are subjects labeled as MF by fitting and MB by estimation. The blue points express subjects that are tagged as MB by fitting and MF by estimation. (*w*=0 for pure MF and *w*=1 for MB one). The median, mean, and standard deviation of estimated values are 0.424, 0.427, and 0.183, respectively wail they were 0.274,0.376, and 0.357 by fitting.

The Daw task has been previously used to investigate the correlation between gaze direction data and *w* [9]. Where 5Param version of the model has been used to extract the *w* value by MLE (see Table 1). We used the *k*-nn estimation to extract the combination weight from their data in the current investigation. The number of trials in [9] has been set to 150, so we make a different database by setting *T* to 150. Fig 10 illustrates the difference between the traditionally "Fitted *w*", used by Konovalov & Krajbich [9], and the "Estimated *w*" which are the result of our estimation method (*k*-nn by $\wp_{\text{sub1}}$ feature space).

To illustrate the differences between MB and MF behavior, Konovalov et al. split subjects into two groups based on the median *w* (0.3). In all analyses, 'model-free' and 'model-based' labels were used for the two groups defined by this median split.

We first check this grouping by the mean value of P-Stay in groups as a behavioral indicator. According to Fig 10, groups changed for some subjects when the estimated *w* was used instead of fitted *w*. The first question arising then is which groups are more consistent with behavioral observation? To answer this question, the stay probability was extracted through different groupings as well as with the group of subjects labeled differently between fitting and estimation (Fig 11).

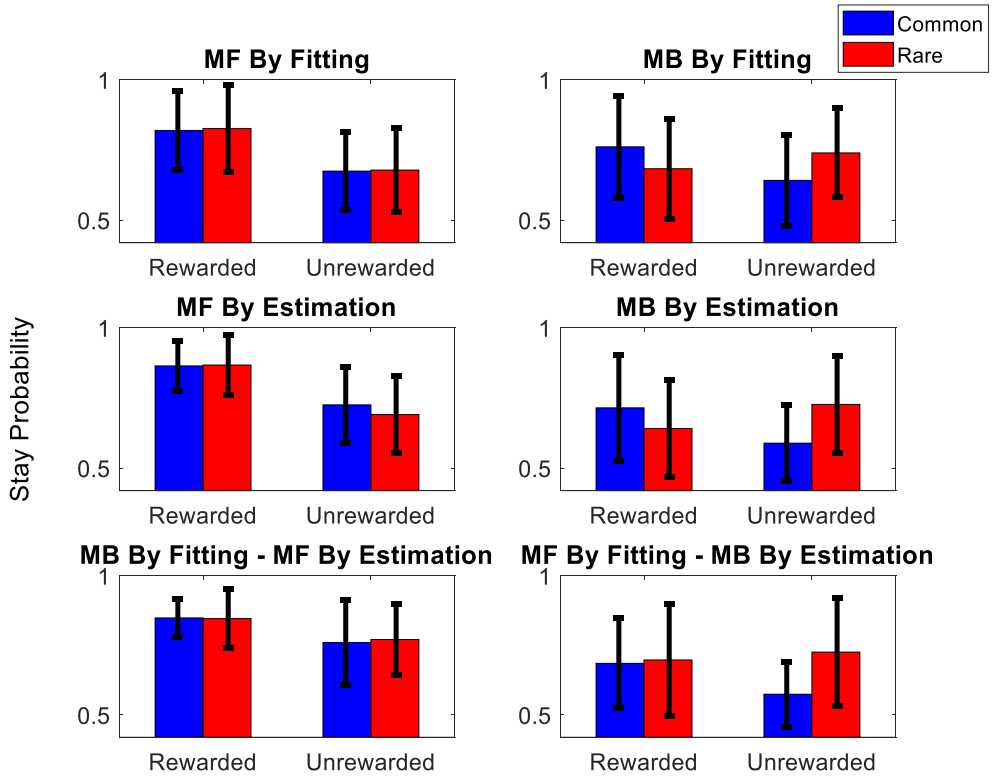

Fig 11 Stay probability for MF and MB groups. The probability is calculated for each subject, and the mean of the calculated value for each group and STD is plotted.

Based on Fig 11, both fitting and estimated groupings are consistent with prior findings. However, while the label assigned by Estimation for groups that are differently labeled is consistent with prior findings on stay probabilities, fitted values show discrepancies. For six subjects that were labeled as MB by fitting and as MF by estimation (blue subjects in Fig 10), the stay probability in trials after rewarded and unrewarded trials is the same in different

transition situations (Common or Rare transition in previous trials); This behavior can be attributed to neglecting the transition (which is the main specification of MF subjects). Therefore, these subjects are a better candidate for MF rather than MB label, and consequently, the estimated *w* is better than that obtained by just fitting.

We check all the analyses given in the first part of the report by Konovalov & Krajbich [9] based on new group labels. While the main analysis results did not change, significant level improvements were observed. An outstanding result of gaze data analyses is the insight into the difference between gaze number distribution of model-based and model-free groups [9].

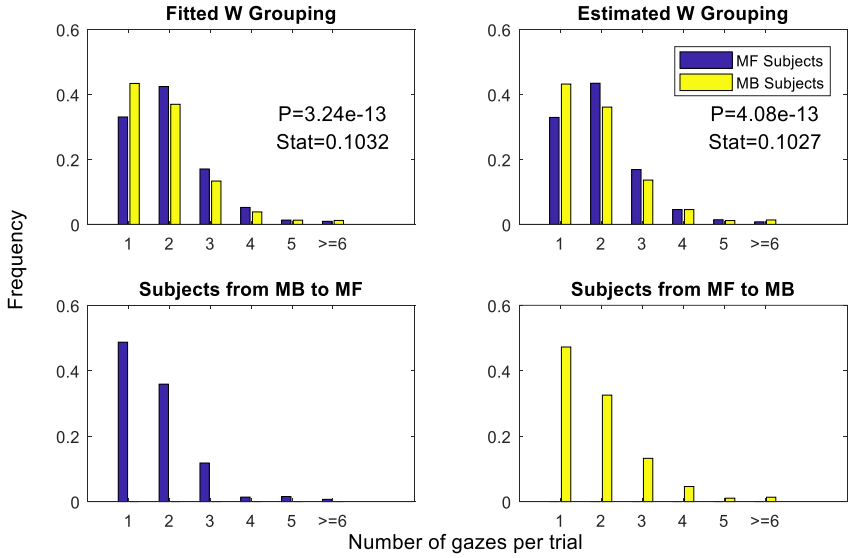

Fig 12 The empirical distribution of gazes' number per trial for different groups. The P-value and statistics for the two-sample Kolmogorov-Smirnov test of the difference, distribution is reported.

Based on Fig 12, the grouping based on estimated *w* makes the difference between MB and MF groups brighter. Model-based subjects were also more likely to look at only one of the symbols before making their first-stage choice (The average for one gaze is 54 vs. 43 by Fitted *w* grouping and 55 vs. 41 by Estimated *w* grouping). MB and MF groups had different distributions for the number of gazes per trial. Statistical test by both groupings demonstrates the significance of the difference (the P-value is 3.24e-13 by fitted *w* grouping and 4.08e-13

by estimated *w* grouping; Kolmogorov-Smirnov test was done over all observations). Moreover, for subjects assumed MF by Fitting but labeled as MB by Estimation, the distribution is closer to the MB group. However, for those subjects that considered MB by fitting but marked as MF by estimation, the distribution is again closer to the MB group.

Moreover, we check the correlation between Estimated w and other behavioral data of subjects. The main study has analyzed some parameters in trial scale and some in subject scales, and there was no correlation between traditionally fitted values and any other available behaviorally meaningful indices. But using the proposed method, we observe that the mean dwell time in middle gazes was strongly correlated with estimated w ($r = 0.4$, $p = 0.008$), while, the traditionally fitted w and the mean dwell time of middle gazes was not correlated ($r=0.08$, $p=0.603$). Fig 13 illustrates this outstanding relationship that was not available by traditionally fitted *w*.

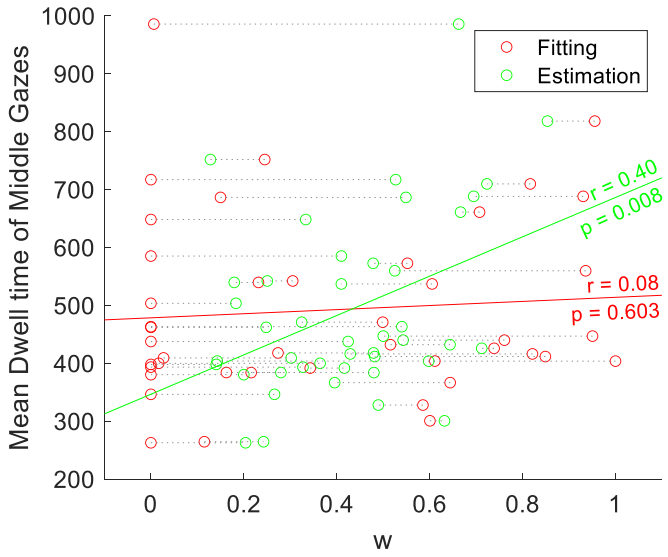

Fig 13 The correlation of the mean dwell time in middle gazes and traditionally fitted *w* (red) and estimated w by proposed method (green). The corresponding correlation coefficients and p-values are reported in the graph.

4.2 Analysis of relationship between learning style and symptom dimension

Gillan et al. recurred the Daw task to investigate the correlation between learning style and compulsive behaviors [35]. Although they use the task with standard-setting, the model that they used for analysis is different. The computational model in their study is a modified of a reparameterization version, introduced by Otto et al. [10]. They showed significant relationships between some psychiatric disorders and the subject's preference towards MB and MF styles. But due to noisy estimation in traditional fitting methods, these relationships are missing for combination weight itself (see below). We believe that recured a more reliable estimation method can revive these correlations in the original model (Section 2.2).

To have a fair comparison, we do the analysis same as their study with the introduced model in section 2.2. and applied both traditional fitting methods and the proposed method. Table 6 reports The correlation coefficient between the self-report questionnaire's total scores and combination weight (w) extracted by the proposed method and traditional fitting methods and corresponding p-values.

As a control for regression analysis, Gillan et.al, used age, IQ, and gender, which has been previously reported to covary with goal-directed behavior [35,46,47]. In line with this study, the extracted combination weight by all three k-nn methods have significant relationships with age, IQ, and Gender but, there is only a relationship between age and the combination-weight fitted by MLE (look at Table 6 for more details). It means that the traditional fitting method extracts the combination weight that is not consistent with other analyses.

Based on regression analysis in Gillan et al. study, there was a significant inverse association between scores on the impulsivity questionnaire and goal-directed behavior (β=-0.034, SE=0.01, p=0.007). Two k-nn methods also replicate this association, but the traditional fitting

methods did not. Also, eating disorders and goal-directed behavior are inversely associated based on regression analysis (β =-0.037, SE=0.01, p<0.001), and the k-nn ($\wp$Sub1) also replicate this association. Alcohol addiction and goal-directed deficits are associated based on regression analysis (β =-0.025, SE=0.01, p=0.029), and the MAP and the k-nn ($\wp$Sub2) method replicate this association.

Gillan et.al. introduced three factors for more analysis, and we also analyzed the correlation between these factors and extracted combination weight by different methods. Based on the regression analysis, there is a significant association between the factor 2 or 'Compulsive Behavior and Intrusive Thought' and goal-directed behavior (β =-0.046, SE=0.01, p<0.001), which the proposed method also replicated it but the traditional fitting methods missed this relationship. Moreover, there were no significant effects of Factor 1 (β =-0.001, SE=0.01, p=0.92) or Factor 3 (β =0.013, SE=0.01, p=0.24) based on both regression analyses and the proposed method, but the traditional fitting method report an association.

In sum, the proposed method could replicate some relationships between goal-directed behavior and some psychiatric disorders, which traditional fitting methods missed this relationship. It can be due to noise reduction in the proposed method relative to the traditional fitting methods. Note that Gillan et al show these relationships by regression and another model. So we can claim that this estimation method is more reliable than traditional methods in finding clinically relevant relationships.

Table 6. Correlation between self-report questionnaire total score and combination weight (r^2 (p-value))

	k-nn ($\wp$)	k-nn ($\wp_{Sub1}$)	k-nn ($\wp_{Sub2}$)	MLE	MAP
Age	4.52e-03 (0.0114)*	4.89e-03 (0.0086)**	2.39e-02 (5.2e-09)**	3.43e-03 (0.0277)*	1.02e-04 (0.7039)
IQ	3.09e-02 (2.8e-11)**	2.94e-02 (8.7e-11)**	5.91e-02 (1.9e-20)**	1.29e-03 (0.1780)	2.80e-04 (0.5294)
Gender	7.60e-03 (0.0010)**	7.38e-03 (0.0012)**	1.36e-02 (1.1e-05)**	8.19e-04 (0.2822)	4.12e-04 (0.4456)

Impulsivity	6.10e-03 (0.0033)**	5.08e-03 (0.0074)**	6.04e-04 (0.3561)	1.15e-03 (0.2031)	1.45e-03 (0.1519)
Eating disorders	1.64e-03 (0.1284)	3.46e-03 (0.0271)*	2.56e-03 (0.0574)	1.85e-04 (0.6099)	1.33e-03 (0.1700)
Alcohol addiction	6.60e-04 (0.3347)	8.05e-05 (0.7361)	3.68e-03 (0.0227)*	2.48e-03 (0.0614)	5.06e-03 (0.0075)**
OCD	2.67e-04 (0.5398)	4.57e-04 (0.4219)	1.47e-04 (0.6488)	2.41e-04 (0.5601)	1.09e-03 (0.2147)
Schizotypy	8.38e-04 (0.2769)	4.62e-04 (0.4193)	5.43e-06 (0.9303)	1.35e-08 (0.9965)	2.26e-03 (0.0741)
Depression	2.59e-04 (0.5456)	1.18e-04 (0.6838)	1.08e-04 (0.6966)	1.60e-04 (0.6350)	1.37e-03 (0.1637)
Trait Anxiety	1.81e-06 (0.9596)	4.94e-05 (0.7918)	3.85e-04 (0.4610)	3.93e-05 (0.8139)	1.08e-03 (0.2160)
Apathy	5.56e-04 (0.3758)	2.28e-04 (0.5705)	1.48e-04 (0.6483)	9.92e-04 (0.2368)	3.42e-04 (0.4870)
Social anxiety	8.99e-05 (0.7218)	8.01e-05 (0.7368)	7.14e-05 (0.7510)	2.58e-06 (0.9519)	2.36e-03 (0.0680)
Factor1	2.08e-06 (0.9568)	7.10e-06 (0.9203)	1.41e-03 (0.1576)	1.51e-04 (0.6446)	2.62e-03 (0.0543)
Factor2	2.84e-03 (0.0453)**	2.82e-03 (0.0460)**	1.83e-03 (0.1084)	1.85e-05 (0.8715)	5.79e-04 (0.3663)
Factor3	5.25e-07 (0.9783)	7.84e-08 (0.9916)	5.21e-05 (0.7863)	5.33e-05 (0.7840)	3.13e-03 (0.0355)*

5 Discussion and conclusion

The extraction of MB and MF learning balance is a necessary step in the transition of reinforcement learning modeling to mathematical psychology. The two-step task of Daw et al. that was designed to disassociate MB and MF learning styles was recently used widely. We study the precision of extracting the subject's preference towards MB and MF styles by this task. To do this, we used nine different versions of the primary model, which are the nested model of the most complex one. To have a performance measure, we observe the simulated model behavior while performing the Daw task, and then the combination weight (w) is extracted from the observed behavior. Because we know the agents' parameters, the estimation

error can be calculated. Our analysis specified that complex models over-fit to the observation and simple models with wrong assumptions result in higher errors. Moreover, when prior knowledge was not assumed for the fitted parameters, the fitted values mostly stick to the extremes of the parameter range. Such problems in model fitting make the fitted parameters unreliable.

We also investigated the effect of learning rate (α) and Boltzmann inverse temperature (β) of agents on model fitting error. The α controls the effectiveness of the new trial in comparison with the previous estimation. Low α values, from an agent point of view, mean that previous estimation is precise enough for decision making. Hence, the new observation for rewarded or unrewarded action makes slight changes in the evaluation. This small change results in the same behavior on MB and MF systems.

Boltzmann inverse temperature, on the other hand, controls the exploration-exploitation tradeoff. The low β values result in a similar choice probability for actions regardless of their values, which means more exploration. In this case, the effect of actions' values that were calculated by either MB or MF systems decreases and marginally (β became zero) is ignored. So, it is expected that the explorative subject has slight information about the preference of MB or MF system, and extracting the subject's preference towards MB and MF styles will be more difficult by any estimation. High β values show that even slightly higher values of action, make them more preferred choices, which is an indication of the exploitative behavior. For higher β values, either little or huge differences in action-values have the same effects on the observer behavior.

From the behavioral data, besides the traditional model fitting, some statistical indices were extracted and used for investigating the cognitive studies. We propose to fuse these two types of information by using k-nn as a simple learning method. Also, just behavioral information

can be used to estimate the parameter instead of fitting the model. We use 39 features (including fitting based features) to generate the *k*-nn dataset and then by forward selection and elimination of fitting based features two different subspace of feature extract. Feature selection can improve the *k*-nn performance, and eliminating the fitted based features reduce both computational load and noise effect. The best performance, which is mean absolute error over different observations, was reached by *k*-nn. Both bias and variance of error were proven to be reduced by *k*-nn learning compared to model fitting. The analysis also specifies that the *k*-nn method is more stable in the presence of decision noise, especially by excluding all fitting based features.

The proposed method is advantageous due to its lower error for extreme cases. Such extreme cases may be prevalent in clinical trials and psychiatric conditions and will make the proposed method to have superior performance over just model-fitting approaches. MAP estimation is better than MLE in extreme values due to using a prior, *k*-nn method works very better than MAP too. The mentioned improvements will enhance the applicability of the Daw task for computational psychiatry purposes.

We showed that using the proposed method can help to increase the statistical power in analyzing the relation between parameters such as the gaze distribution to habitual and goal-directed behavior. It was proven that consideration of behavioral parameters in the estimation of combination weight (in addition to fitting), improves the consistency of behavior and subjects grouping and so other conclusions from this grouping, can be more precise.

Using the proposed method on clinical subjects has extracted some relationships between disorders and habitual vs. goal-directed behavior axis which were missed by traditional fitting methods. These relationships were validated by a reparametrized model in Gillan et al study [35]. Because adding some noise to one variable can destroy the correlation coefficient between that

variable and other measures, it seems that some correlation coefficient has lost their significance due to noisy estimation of combination weight, and the proposed method was more successful due to reduction of this noise. It worth noting that though the proposed method was successful in extracting most relationships of this study, some relationships were missing in the proposed method too. For example, there was an association between OCD and goal-directed behavior based on regression analysis but none of w extraction methods reflect that.

Note that any model fitting tries to minimize an objective function to extract the behavior under different assumptions. The MLE maximizes the likelihood function, while the extracted parameter by *k*-nn will not maximize the likelihood, although the estimation error in *k*-nn is lower. The flow of probabilities in reinforcement agent decisions causes that a specific parameter does not guarantee maximum likelihood while another parameter exists which satisfies the maximizes likelihood criterion. Although MLE can theoretically obtain the Cramer-Rao Lower Band, the above statement is the cause that learning reaches better estimations rather than MLE.

The proposed method can be considered as a maximum likelihood estimation using simulation-based estimation. Such a method not only uses trial-by-trial observations of the behavior but also uses global observation such as stay probabilities in random variable space and tries to maximize the likelihood of observing all the mentioned behaviors together. For large sample sizes, MLE and *k*-nn methods may converge to the same estimation error. For limited sample sizes, however, *k*-nn has shown more reliability and avoids overfitting, and is considered a better option in a typical experimental condition.

In sum, our proposed method can enhance the estimation of the model-based and model-free combination weight. This improvement is due to using behavioral indices from the data

that make the estimation more robust. This robust estimation can facilitate the handling of similar paradigms in clinical applications and help in the diagnosis of psychiatric disorders.

Acknowledgment

We would like to express our gratitude to Arkady Konovalov and Ian Krajbichv for sharing the gaze task data with us as well as Gillan et.al. Also We would like to thank the reviewers for their thoughtful comments and efforts towards improving our manuscript.

Reference

1. Wanjerkhede, S. M., Bapi, R. S. & Mytri, V. D. Reinforcement learning and dopamine in the striatum: A modeling perspective. *Neurocomputing* **138**, 27–40 (2014).
2. Dolan, R. J. & Dayan, P. Goals and habits in the brain. *Neuron* **80**, 312–325 (2013).
3. Keramati, M., Smittenaar, P., Dolan, R. J. & Dayan, P. Adaptive integration of habits into depth-limited planning defines a habitual-goal–directed spectrum. *Proc. Natl. Acad. Sci.* **113**, 12868–12873 (2016).
4. Kool, W., Cushman, F. A. & Gershman, S. J. When Does Model-Based Control Pay Off? *PLoS Comput. Biol.* **12**, 1–34 (2016).
5. Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–1711 (2005).
6. Lucantonio, F., Caprioli, D. & Schoenbaum, G. Transition from 'model-based' to 'model-free' behavioral control in addiction: Involvement of the orbitofrontal cortex and dorsolateral striatum. *Neuropharmacology* **76**, 407–415 (2014).
7. Toyama, A., Katahira, K. & Ohira, H. A simple computational algorithm of model-based choice preference. *Cogn. Affect. Behav. Neurosci.* **17**, 764–783 (2017).
8. Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**, 1204–1215 (2011).
9. Konovalov, A. & Krajbich, I. Gaze data reveal distinct choice processes underlying model-based and model-free reinforcement learning. *Nat. Commun.* **7**, 12438 (2016).
10. Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A. & Daw, N. D. Working-memory capacity protects model-based learning from stress. *Proc. Natl. Acad. Sci.* **110**, 20941–20946 (2013).
11. Ahn, W. Y. & Busemeyer, J. R. Challenges and promises for translating computational tools into clinical practice. *Curr. Opin. Behav. Sci.* **11**, 1–7 (2016).
12. Montague, P. R., Dolan, R. J., Friston, K. J. & Dayan, P. Computational psychiatry. *Trends Cogn. Sci.* **16**, 72–80 (2013).
13. Gillan, C. M. & Robbins, T. W. Goal-directed learning and obsessive–compulsive disorder. *Philos. Trans. R. Soc. B Biol. Sci.* **369**, 20130475 (2014).
14. Lucantonio, F., Caprioli, D. & Schoenbaum2, G. Transition from 'model-based' to 'model-free' behavioral control in addiction: involvement of the orbitofrontal cortex and dorsolateral striatum. *Bone* **23**, 1–7 (2015).
15. Everitt, B. J. & Robbins, T. W. Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. *Nat. Neurosci.* **8**, 1481–1489 (2005).

16. Gillan, C. M. *et al.* Disruption in the balance between goal-directed behavior and habit learning in obsessive-compulsive disorder. *Am. J. Psychiatry* **168**, 718–726 (2011).
17. Gillan, C. M. & Daw, N. D. *Taking Psychiatry Research Online Claire*. *Neuron* vol. 91 19–23 (Cell Press, 2016).
18. Voon, V. *et al.* Disorders of compulsivity: a common bias towards learning habits. *Mol. Psychiatry* **20**, 345–352 (2015).
19. Toyama, A., Katahira, K. & Ohira, H. Biases in estimating the balance between model-free and model-based learning systems due to model misspecification. *J. Math. Psychol.* **91**, 88–102 (2019).
20. de Wit, S., Barker, R. A., Dickinson, A. D. & Cools, R. Habitual versus Goal-directed Action Control in Parkinson Disease. *J. Cogn. Neurosci.* **23**, 1218–1229 (2011).
21. Culbreth, A. J., Westbrook, A., Daw, N. D., Botvinick, M. & Barch, D. M. Reduced model-based decision-making in schizophrenia. *J. Abnorm. Psychol.* **125**, 777–787 (2016).
22. Foerde, K. What are habits and do they depend on the striatum? A view from the study of neuropsychological populations. *Curr. Opin. Behav. Sci.* **20**, 17–24 (2018).
23. Daw, N. D. Of goals and habits. *Proc. Natl. Acad. Sci.* **112**, 13749–13750 (2015).
24. Smittenaar, P., FitzGerald, T. H. B., Romei, V., Wright, N. D. & Dolan, R. J. Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans. *Neuron* **80**, 914–9 (2013).
25. Dezfouli, A. & Balleine, B. W. Actions, Action Sequences and Habits: Evidence That Goal-Directed and Habitual Action Control Are Hierarchically Organized. *PLoS Comput. Biol.* **9**, (2013).
26. Sebold, M. *et al.* Model-based and model-free decisions in alcohol dependence. *Neuropsychobiology* **70**, 122–131 (2014).
27. Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D. & Daw, N. D. Model-based choices involve prospective neural activity. *Nat. Neurosci.* **18**, 767–772 (2015).
28. Cushman, F. & Morris, A. Habitual control of goal selection in humans. *Proc. Natl. Acad. Sci.* **112**, 13817–13822 (2015).
29. Kool, W., Gershman, S. J. & Cushman, F. A. Cost-Benefit Arbitration Between Multiple Reinforcement-Learning Systems. *Psychol. Sci.* **28**, 1321–1333 (2017).
30. Morris, L. S., Baek, K. & Voon, V. Distinct cortico-striatal connections with subthalamic nucleus underlie facets of compulsivity. *Cortex* **88**, 143–150 (2017).
31. Gillan, C. M., Otto, A. R., Phelps, E. A. & Daw, N. D. Model-based learning protects against forming habits. *Cogn. Affect. Behav. Neurosci.* **15**, 523–536 (2015).
32. Guitart-Masip, M. *et al.* Differential, but not opponent, effects of l-DOPA and citalopram on action learning with reward and punishment. *Psychopharmacology (Berl).* **231**, 955–966 (2014).
33. Kroemer, N. B. *et al.* L-DOPA reduces model-free control of behavior by attenuating the transfer of value to action. *Neuroimage* **186**, 113–125 (2019).
34. Sutton, R. S. & Barto, A. G. *Introduction to reinforcement learning*. (MIT Press, 1998).
35. Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A. & Daw, N. D. Characterizing a psychiatric symptom dimension related to deficits in goaldirected control. *Elife* **5**, 1–24 (2016).
36. Shahar, N. *et al.* Improving the reliability of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-diffusion modeling. *PLOS Comput. Biol.* **15**, e1006803 (2019).
37. Ward, M. D., Carolina, N. & Ahlquist, J. S. Maximum Likelihood for Social Sciences

Strategies for Analysis Chapter 1 Introduction to Maximum Likelihood. (2012).
38. Breidt, F. J. & Opsomer, J. D. Model-assisted survey estimation with modern prediction techniques. *Stat. Sci.* **32**, 190–205 (2017).
39. Li, Z., Liu, G. & Li, Q. Nonparametric Knn estimation with monotone constraints. *Econom. Rev.* 1–19 (2017).
40. Silverman, B. W. & Jones, M. C. E. Fix and J.L. Hodges (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951). *Int. Stat. Rev. / Rev. Int. Stat.* **57**, 233 (1989).
41. Cover, T. M. Estimation by the Nearest Neighbor Rule. *IEEE Trans. Inf. Theory* **14**, 50–55 (1968).
42. Dudani, S. A. The Distance-Weighted k-Nearest-Neighbor Rule. *IEEE Trans. Syst. Man Cybern.* **SMC-6**, 325–327 (1976).
43. Collins, A. G. E., Albrecht, M. A., Waltz, J. A., Gold, J. M. & Frank, M. J. Interactions Among Working Memory, Reinforcement Learning, and Effort in Value-Based Choice: A New Paradigm and Selective Deficits in Schizophrenia. *Biol. Psychiatry* **82**, 431–439 (2017).
44. Miller, K. J. *et al.* Identifying Model-Based and Model-Free Patterns in Behavior on Multi-Step Tasks. *bioRxiv* 096339 (2016) doi:10.1101/096339.
45. Wichmann, F. A. & Hill, N. J. The psychometric function: I. Fitting, sampling, and goodness of fit. *Percept. Psychophys.* **63**, 1293–1313 (2001).
46. Eppinger, B., Walter, M., Heekeren, H. R. & Li, S. C. Of goals and habits: Age-related and individual differences in goal-directed decision-making. *Front. Neurosci.* **7**, 1–14 (2013).
47. Schad, D. J. *et al.* Processing speed enhances model-based over model-free reinforcement learning in the presence of high working memory functioning. *Front. Psychol.* **5**, 1–10 (2014).