Title: Better than Maximum Likelihood Estimation of Model-based and Model-free Learning Styles

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

1. *School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran.*

***Corresponding Author**: Abdol-Hossein Vahabie, School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran. Email: h.vahabie@ut.ac.ir

To appear in: **Basic and Clinical Neuroscience**

Received date: 2023/09/30

Revised date: 2024/07/28

Accepted date: 2024/10/06

Please cite this article as:

Yazdani, S., Vahabie, A.H., Nadjar-Araabi, B., Nili Ahmadabadi, M. (In Press). Better than Maximum Likelihood Estimation of Model-based and Model-free Learning Styles. Basic and Clinical Neuroscience. Just Accepted publication Jul. 10, 2024. Doi: http://dx.doi.org/10.32598/bcn.2024.5883.1
DOI: http://dx.doi.org/10.32598/bcn.2024.5883.1

Abstract:

Various decision-making systems work together to shape human behavior. Goal-directed and habitual systems are the two most important systems studied by reinforcement learning (RL) through model-based (MB) and model-free (MF) learning styles, respectively. Human behavior resembles the combination of these two decision-making paradigms, achieved by the weighted sum of the action values of the two styles in an RL framework. The weighting parameter is often extracted by the maximum likelihood (ML) or maximum a-posteriori (MAP) estimation method. In this study, we employ RL agents that use a combination of MB and MF decision-making to perform the well-known Daw two-stage task. ML and MAP methods result in less reliable estimates of the weighting parameter, where a large bias toward extreme values is often observed. We propose the *k*-nearest neighbor as an alternative nonparametric estimate to improve the estimation error, where we devise a set of 20 features extracted from the behavior of the RL agent. Simulated experiments examine the proposed method. Our method reduces the bias and variance of the estimation error based on the obtained results. Human behavior data from previous studies is investigated as well. The proposed method results in predicting indices such as age, gender, IQ, the dwell time of gaze, and psychiatric disorder indices which are missed by the traditional method. 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 maximum likelihood and maximum a-posteriori estimations; Behavioral observation analysis; Behavioral parameter estimation

1 Introduction

Multiple cognitive systems are thought to control human decision-making. Most decisions and learning occur during a person's lifespan as a result of habitual and goal-directed systems (Dolan & Dayan, 2013; Wanjerkhede et al., 2014). The habitual system creates habits and automatic decisions, whereas the goal-directed system is primarily concerned with planning. Researchers studying reinforcement learning (RL) assign habitual and goal-directed systems to model-based (MB) and model-free (MF) learning styles. The only distinction between the MB and MF styles is in the evaluation of state-action. In MB learning, an environmental model is used to evaluate each decision in the current state; in MF learning, action-values update without the use of any explicit environment model, and the value of each action in each state is learnt via trial and error.

Previous studies found that people employ a combination of MB and MF learnings to direct their behavior during learning tasks. Several research support the notion that the hybrid model is an effective subject description(Daw et al., 2005; Dolan & Dayan, 2013; Gijsen et al., 2022; Keramati et al., 2016; Kool et al., 2016; Lucantonio et al., 2014; Toyama et al., 2017). The combination weight (*w*) is the parameter that affects the subject's preference for MB in this model, as explained in the Supplementary 1(Daw et al., 2011).

Computational models can assist extract various cognitive components driving maladaptive behavior, and the model parameters associated with those components can be utilized to investigate the potential sources of cognitive deficiencies (Ahn & Busemeyer, 2016). One of the elements that can be used to analyze, diagnose, and evaluate the efficacy of therapies for psychiatric diseases is the parameter that determines the subject's preference for the MB/MF style (Montague et al., 2013). In a two-stage task that Daw and his associates proposed, the reward probability in the second stage fluctuates over time and the first stage transition is probabilistic(Daw et al. 2011). As a result, the MB and MF styles behave differently. This task is frequently used by researchers to determine how much the participant prefers MB and MF styles. (Daw, 2015; Doll et al., 2015; Feher da Silva & Hare, 2020; Foerde, 2018; Gillan et al., 2015; Miller et al., 2022; Morris et al., 2017; Otto et al., 2013; Smittenaar et al., 2013).

Considering changes in the subject’s preference toward MB (*w*) due to pharmacological or cognitive manipulations or neuropsychiatric conditions will provide important insights for clinical research. For example, Over-reliance on the MF style could lead to inflexible decisions in addiction and compulsion (Everitt & Robbins, 2005; Gillan & Robbins, 2014; Lucantonio et al., 2014). Some studies show that patients with obsessive-compulsive disorder (OCD) prefer the MF learning style more than MB (Gillan et al., 2011; Gillan & Daw, 2016; Toyama et al., 2019; Voon et al., 2015). Wit and colleague, show that mild Parkinson’s disease has led to impaired MF habit formation (Wit et al., 2011). Also, Culberth and colleague, show that in schizophrenic patients, MB behavior is reduced (Culbreth et al., 2016). On a broader view, there is a growing consensus that computational modeling can be constructive in understanding 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, and low sample size, especially for extreme values.

Traditionally, researchers estimated the model parameters, such as the subject's preference for MB(*w*), by fitting the model to their observations using maximum likelihood (ML) or maximum a posteriori (MAP). The best objective function for model fitting is likelihood when there is no other information than behavioral observation. The foundation of ML is the notion that a specific collection of parameters has a greater likelihood of being responsible for the observed data. ML is widely applied in the behavioral sciences (Ward et al., 2012). In addition, if we are aware of any prior knowledge of parameters, we employ the MAP approach.

According to the analysis, the precision of the estimated combination weight based on conventional model fitting is subpar. Traditional methods' precision is affected by factors such as the nature of the task, the model, noise, the fitting method, and the limited number of observations. Our simulations demonstrate that the conventional estimation technique is biased toward the MF style, particularly when the other model parameters are outside the acceptable range. In model fitting, the estimation of the combination weight is more inaccurate when the learning rate or temperature are low or high, respectively (see supplementary 2 for details).

In the present research, we propose that using a data-driven learning method in addition to the traditional fitting methods can improve estimation precision and reliability. This research employs the *k*-nearest neighbor (*k*-nn) algorithm as a straightforward learning technique (see supplementary 3 for more details). Other learning algorithms, such as deep neural networks, can serve the same function. Although this study focuses on the observation of action selection, the estimator can be made more precise by incorporating other measurable parameters, such as confidence level or response time (Shahar et al., 2019).

In this study, we attempt to improve the estimation of a model's parameter based on behavioral observation as compared to the common traditional method. Although we analyze the effect of some nested models on parameter estimation error, we do not investigate which model is superior in other ways, such as predicting human behavior. This study did not investigate alternative models, such as the Gijsen model (Gijsen et al., 2022). Some studies use the reparameterization method (alternative models with different free parameters) or other combinations of reparameterization (Gillan et al., 2016; Toyama et al., 2019). In addition, some studies utilize the response time of a model that is unavailable for our simulation and was not incorporated into the model (Shahar et al., 2019). Although a subject's preference for a particular style can change over time, we will assume that it remains constant for the duration of the task.

Section 2 discusses the basic model architecture (Section 2.1) and the *k*-nn estimator (Section 2.2). 2.3 and 2.4 explain implementation. Section 3 sets the *k* parameter (section 3.1) and analyzes the proposed method's result (section 3.2). Section 3.3 analyzes the *w* extraction in a noisy model. Section 4 shows the *k*-nn method's experimental performance and its advantages. The conclusion discusses the proposed method and summarizes the results (section 5).

2 Method

In this study, we compare the results of determine preference for MB (*w*) using traditional method and proposed method for humans and simulated agents. In simulation and the ML and MAP methods, we employ the Daw et al. model (see supplementary 1 for details). During the training and testing phases, the behavioral data is derived from simulation, whereas during the recall phase, it is derived from actual human behavior. Fig 1 depicts the proposed method overall.

In this paper, in addition to the estimated values of the parameter by ML or MAP, we use global information, including behavior statistics and indices, to extract the subject's preference for MB (*w*) more precisely. In the proposed method, the *k*-nn estimator (also known as the *k*-nn regressor) is employed as a learning system to extract *w* from behavior. *k*-nn is a supervised, nonparametric learning method that has been widely adopted as an accurate point estimator (Li et al., 2017). The *w* parameter is estimated by *k*-nn using a set of labeled feature vectors. To train the *k*-nn, we employ simulations of RL agents, and the dataset is populated with features derived from observations labeled by the agent's combination weight parameter (w_o).

Since we know the parameters of the RL agent during the testing phase, the estimation error can be calculated. We illustrate the error distribution using the mean absolute error (MAE) as a point estimator of the error and the standard deviation (STD). The simulations contain a sufficient number of agents for reliable results so following the statistical tests results were not reported for simulation data.

In the current study, the objective functions are minimized by the interior-point optimization algorithm, and ten random starting points are used to maximize the probability of global optimization for ML and MAP. All analyses and optimizations have been implemented in MATLAB 2021b and are accessible via Dataverse repository: https://doi.org/10.7910/DVN/PSEFZF.

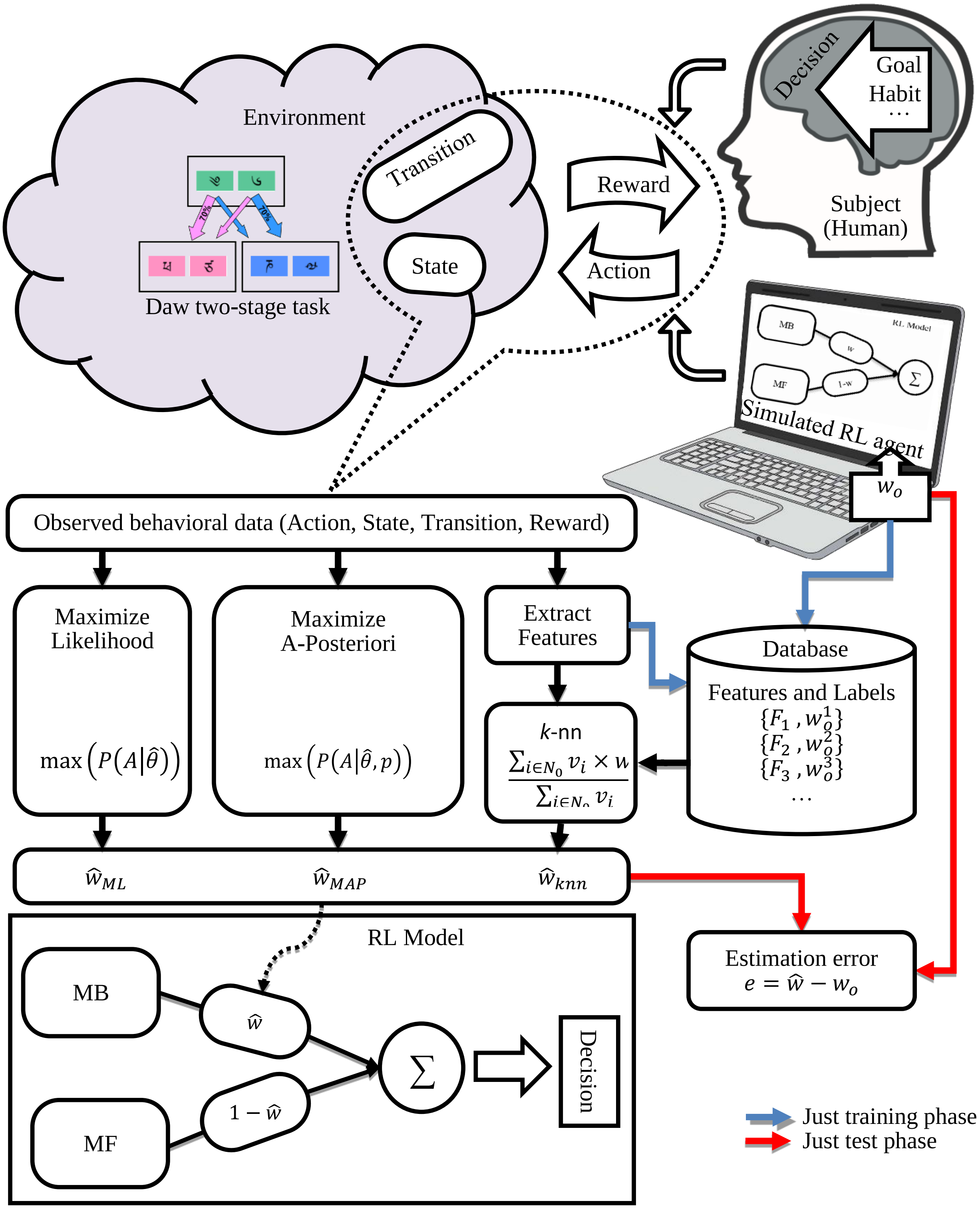

Fig 1. The flow of information in ML, MAP, and the proposed method (*k*-nn) in the training and test phase, the simulated RL agent performs the task, and in the recall phase, observed data from human behavior is used.

2.1 Computational Model

Daw et al. proposed a computational model predicated on the notion that subjects utilize both MB and MF learning styles, with the values being linearly combined. They suggest using the SARSA-λ algorithm to extract the MF style value and the Bellman equation to extract the MB style value. Using a linear weighted combination, the net value of an action (a) in a state (s) is computed for each trial (t) (Equation 1).

$$Q_{net}^{t}(s,a) = w \times Q_{MB}^{t}(s,a) + (1-w) \times Q_{MF}^{t}(s,a) \tag{1}$$

The free parameter w represents the subject's MB learning style preference. Then, the value of the same previous action increases by the stickiness parameter(p), and the model extracts the probability of decisions using soft-max. Each trial is updated by incremental learning, which modifies the state-action values (See Supplementary 1 for details). Multiple other researchers have employed this hybrid model as well (Kroemer et al., 2019; Morris et al., 2017).

The Daw model for the task contains seven parameters (DS-λE-SS), but in many studies, some of these parameters are set identically in two stages or are assumed to have a constant value. We extracted nine model versions for analysis using this method. The models and subsets of each version's parameters are detailed in Table **1**.

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

Parameter Name	Combination weight of MB/MF	Learning Rate 1st Stage	Learning Rate 2nd Stage	Inverse Temperature 1st Stage	Inverse Temperature 2nd Stage	Eligibility Trace	Stickiness to repeating the same 1st action	Stickiness to repeating the same 2nd action	Number of parameters
Parameter Symbol Version*	w	α_1	α_2	β_1	β_2	λ	P_1	P_2	
IS-1E-NS	w	α	α	β	β	1	0	0	3
IS-0E-NS	w	α	α	β	β	0	0	0	3
IS-λE-NS	w	α	α	β	β	λ	0	0	4
DS-1E-NS	w	α_1	α_2	β_1	β_2	1	0	0	5
DS-0E-NS	w	α_1	α_2	β_1	β_2	0	0	0	5
DS-λE-NS	w	α_1	α_2	β_1	β_2	λ	0	0	6
DS-λE-SS	w	α_1	α_2	β_1	β_2	λ	P	P	7
DS-λE-1S	w	α_1	α_2	β_1	β_2	λ	P	0	7
DS-λE-DS	w	α_1	α_2	β_1	β_2	λ	P_1	P_2	8

*Version Name: XS-YE-ZS

X
I: identical α and β for both Stages
D: different α and β for Stages

Y
0: No Eligibility
1: Full Eligibility
λ: Grade of Eligibility between 0 to 1

Z
N: No Stickiness to repeating
S: Same Stickiness to repeating in two stages
1: Stickiness to repeating just in 1st stage
D: Different Stickiness to repeating in two stages

2.2 The *k*-nn

The distance-weighted method of the *k*-nn estimator is utilized. Two groups of behavioral observation-derived characteristics are listed in Table 2. There are ten characteristics within each group. The first set is based on the stay-probability, which is calculated by counting the stays in observed behavior, i.e., selecting the same action as the previous trial in the first stage. Numerous studies utilizing the Daw two-stage task employed the conditional stay probability for analysis (Collins et al., 2017; Daw et al., 2011). We employ the stay probability across situations and conditions based on the reward value (either rewarded or unrewarded) and transition frequency (common or uncommon) of previous trials. In addition, the slope of stay probabilities, as indices for MF (equation (2)) and MB (equation (3)) behavior (Miller et al., 2016), was utilized as an additional behavioral indicator in feature space.

$$\mathrm{I}_{MF}^{\mathrm{PStay}} = P\big(S \,\big|\, Re, C\big) + P(S|Re,R) - P(S|Ur,C) - P(S|Ur,R) \tag{2}$$

$$\mathrm{I}_{MB}^{\mathrm{PStay}} = P\big(S \,\big|\, Re, C\big) - P(S|Re,R) - P(S|Ur,C) + P(S|Ur,R) \tag{3}$$

The second group consists of model-parameter-using and model-fitting features. Miller et al. introduced the MB/MF preference indexes, equation (4) and (5), which we employ (Miller et al., 2016).

$$\mathrm{I}_{MF}^{\mathrm{Fit}} = (1 - \widehat{w}_{Fit}) \times \hat{\beta}_1^{Fit} \tag{4}$$

$$\mathrm{I}_{MB}^{\mathrm{Fit}} = \widehat{w}_{Fit} \times \hat{\beta}_1^{Fit} \tag{5}$$

In these equations, $\hat{w}$ and $\hat{\beta}_1$ are the combination weight and inverse temperature of the first stage, respectively, and are derived by fitting the model using ML or MAP. In addition, we add some RL model parameters, such as the combination weight (w) itself, estimated by fitting the model with ML or MAP (see supplementary 3 for details).

2.3 Generated Dataset for *k*-nn

As a supervised learning technique, *k*-nn requires a training dataset with the appropriate labels to perform properly. Therefore, we simulate 80,000 independent RL agents with random parameters and the DS-λE-DS version (see Table 1), and then record their behavioral observations. In This study, we picked all random parameters and MAP prior knowledge according to Table 3.Each simulation includes a series of trials and associated observations, all of which are tagged with the w_o. In addition, 10-fold cross-validation is utilized to tune the hyper-parameter *k*. To eliminate estimator bias at extremes, we augment the training dataset with 10,000 MB and 10,000 MF agents.

Table 2. Features Set

#		Symbol	Description
1	based on stay probability	$P(S\mid Re)$	Stay probability over trials after the Rewarded trial
2		$P(S\mid Ur)$	Stay probability over trials after the Unrewarded trial
3		$P(S\mid C)$	Stay probability over trials after the Common trial
4		$P(S\mid R)$	Stay probability over trials after the Rare trial
5		$P(S\mid Re,C)$	Stay probability over trials after different situations across Rewarded, Unrewarded, Common and Rare of the previous trial.
6		$P(S\mid Re,R)$	
7		$P(S\mid Ur,C)$	
8		$P(S\mid Ur,R)$	
9		I_{MF}^{PStay}	$\mathrm{I}_{MF}^{\mathrm{PStay}} = P(S\mid Re,C) + P(S\mid Re,R) - P(S\mid Ur,C) - P(S\mid Ur,R)$
10		I_{MB}^{PStay}	$\mathrm{I}_{MB}^{\mathrm{PStay}} = P(S\mid Re,C) - P(S\mid Re,R) - P(S\mid Ur,C) + P(S\mid Ur,R)$
11	based on model fitting	I_{MF}^{MLE}	$\mathrm{I}_{MF}^{\mathrm{Fit}} = (1 - \widehat{w}^{Fit}) \times \hat{\beta}_1$ $\mathrm{I}_{MB}^{\mathrm{Fit}} = \widehat{w}^{Fit} \times \hat{\beta}_1$
12		I_{MB}^{MLE}	
13		I_{MF}^{MAP}	
14		I_{MB}^{MAP}	
15		$\widehat{w}^{MLE}$	Parameters Extracted by Model Fitting
16		$\hat{\alpha}_1^{MLE}$	
17		$\hat{\beta}_1^{MLE}$	
18		$\widehat{w}^{MAP}$	
19		$\hat{\alpha}_1^{MAP}$	
20		$\hat{\beta}_1^{MAP}$	

Table 3. Parameters, range, and random values for independent agents.
For simulation and the prior in MAP Method (Beta(.) is the beta distribution)

Parameter Symbol	Description	Min	Max	Probability density
w	MB/MF combination weight	0	1	Uniform(0,1)
$\alpha_1, \alpha_2, \alpha$	1st and 2nd Stage Learning Rate	0	1	Beta(1.2,1.2)
β_1, β_2, β	1st and 2nd Stage Inverse Temperature	1	10	1+9×Beta(1.2,1.2)
λ	Eligibility Trace	0	1	Beta(1.2,1.2)
P_1, P_2, P	1st and 2nd Stage Stickiness to repeating the same action	0	0.2	uniform(0,0.2)

2.4 Model the lapse in Decision-making

It has been demonstrated that incorporating the lapse rate into models for human subjects can increase the quality of fit for numerous psychophysical paradigms (Wichmann & Hill, 2001). This lapse rate is a result of the random and unattended trials in which the participant participated. We add this noise source capability for agents in simulations. Each agent's choice is reversed based on a probability known as the lapse rate or noise level. We simulate the noisy model with varying lapse rates in the interval [0, 0.5].

3 Results

In this section, we will begin by setting the *k* parameter of the *k*-nn algorithm, and then we will apply the statistics and visualizations necessary to show how well the suggested technique performs. According to the findings of the analyses, both the variance and the bias of the estimation decreased.

3.1 *k*-nn Parameter

The value of *k* affects the effectiveness of *k*-nn. The value of *k* determines the localization and generalization of *k*-nn, and a trade-off between these two factors is required for optimal performance.

To achieve the best *k*-nn performance, we optimize the k value using exhaustive search to minimize MAE. Experimentally, the MAE is nearly constant when k is greater than 40 and less than 100. The MAE varies minimally within the range of 0.1857 to 0.1862 for these values of k, but the optimal value of k is 69, and we use this value in all situations.

Feature selection can improve *k*-nn's performance. We used both unsupervised (analyze feature correlation) and supervised (Backward elimination method) feature selection on the data set, but the performance improvement was minor, so we ignored the feature selection (see supplementary 3 for details).

As mentioned previously, Table 2 contains two feature groups. The first group of features is calculated based on the stay probability. The second group of features requires the fitting procedure, which is complicated by computational load, model selection, and optimization algorithm. To adjust the proposed method for some practical applications in which the mentioned factors restrict the use of fitted parameters, we can disregard the second group of features and, as a result, decrease the method's performance (although in some cases, like having not a good model or noisy observation, this neglecting can improve the performance). We utilized *k*-nn in two distinct circumstances based on the available data and analytics:

1- $\wp_1$: Just 1st Group available (Features from model fitting are excluded)

2- $\wp_{1+2}$:All features will be computed (needs model fitting, i.e., ML and MAP estimation).

3.2 Performance

Fig 2 depicts the scattering of the extracted w by the *k*-nn estimator and traditional model fitting relative to the corresponding value of agents. To have a clear view, we have divided it into five areas. We are aware that the exact combination weight (w_o) cannot be determined due to limited data, so a small error is acceptable. We assume an error of less than 0.1 is tolerable. Grouping subjects by learning style is an application of extracting the *w*. Therefore, if an error in *w* extraction results in the incorrect subjevct label, the error is considerable. The areas that were not altered by the dominant strategy are considered slight errors. Those zones without a dominant strategy (0.45<w_o<0.55 or 0.45< $\hat{w}$<0.55) were assumed to be transition areas. The fifth region is extreme value of $\hat{w}$. According to Fig 2, the results of *k*-nn *w* estimations using $\wp_{1+2}$ features have the highest proportion of the tolerable area. Moreover, the ML and MAP are biased towards the MF style, but *k*-nn methods resolve this issue. In addition, the scatter plot demonstrates that *k*-nn addresses the most problematic aspect of traditional fitting methods, which is the adherence to extreme values.

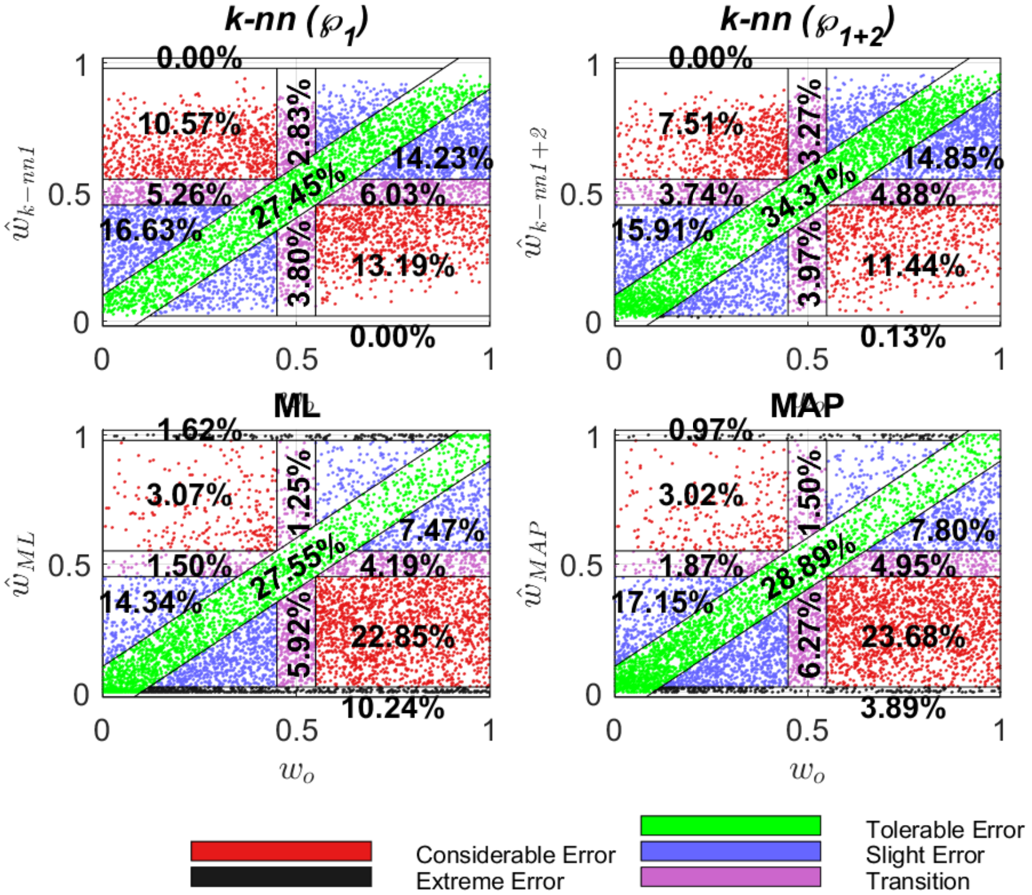

Fig 2. The difference in performance. The horizontal axis represents the agent's combination weight (w_o), and the vertical axis represents the estimated weight ($\hat{w}$). We simulate 10,000 agents performing the Daw task using the DS-λE-DS model and random parameters. We fit all model versions to observation by ML and MAP fitting methods, and then the best version was selected based on the AIC. The points with a low inaccuracy (below 0.1) are assumed to be tolerable and are demonstrated in green. The considerable error area (red points) corresponds to instances in which the dominant style changed between MB and MF. Slight errors are the blue points that indicate that the dominating strategy has not changed. Those regions that lacked a dominant strategy ($0.45<w_o<0.55$ or $0.45<\hat{w}<0.55$) were presumed to be transition area (magenta color). The top and bottom regions are those spots where the extracted w adheres to the extreme and is dispersed in black. Distribution of the points, clarified by percentage, in any area.

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. . Fig 3 illustrates distribution of error, the difference between estimated and true values.

Fig 3 demonstrates that the *k*-nn technique reduces both bias and variance of error. For the *k*-nn approach, the tail of the distribution consists of lower values. The STD of errors verify that the k-nn error variance is better than traditional methods (Table 4). In contrast, the chance of tolerable error (errors between -0.1 and 0.1) is greater for *k*-nn approaches than for fitting methods. In addition, Table 4's presentation of the MAE and correlation coefficient demonstrates that the *k*-nn estimation reduce bias and error variance. Extreme errors are substantially more in ML and MAP than in *k*-nn-based algorithms. Since extreme values for the subject's preference for MB and MF styles are possible under clinical situations, these regions are significant. *k*-nn approaches correct these errors and make the clinical trial model more robust. In accordance with Toyama's work, the skewness of the error in Fig 3 indicates a bias toward MF (Toyama et al., 2019).

Table 4. MAE, STD, and R (pearson r) of *w* estimation error by *k*-nn method and model fitting.

Estimation Method	MAE	STD	R
k-nn ($\wp_1$)	0.2257	0.1665	0.4284
k-nn ($\wp_{1+2}$)	0.1962	0.1592	0.5929
ML	0.2699	0.2207	0.4509
MAP	0.2547	0.2116	0.4608

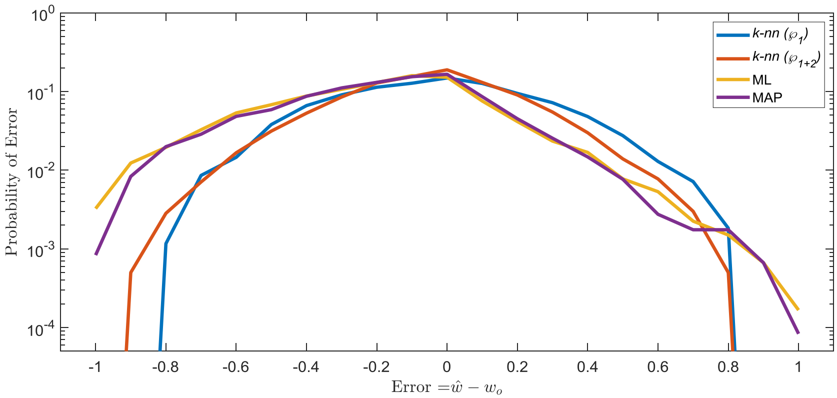

Fig 3. The error distribution for various w extraction. 10,000 independent agents performed the Daw task using the DS-λE-DS model version and random parameters to conduct this analysis. After extracting *w* using each of the previous mentioned techniques, the estimation error (extracted value minus the true value) is computed. For fitting method, *w* is extracted by comparing the AIC of the model version. The output of *k*-nn was computed in the distinct $\wp_1$ and $\wp_{1+2}$ feature spaces. The confidence interval is verry close to the results.

3.3 Lapse in Decision-making

The potential of erroneously selecting the desired option due to attentional lapses or other issues is a real concern in parameter estimation for human data. When considering the effectiveness and applicability of an estimation technique, we should consider its resilience in the face of lapse rates.

We simulate the model with different lapse rates to see what happens when people make mistakes. Fig 4 shows how the knn estimation method is different from traditional ways of fitting.

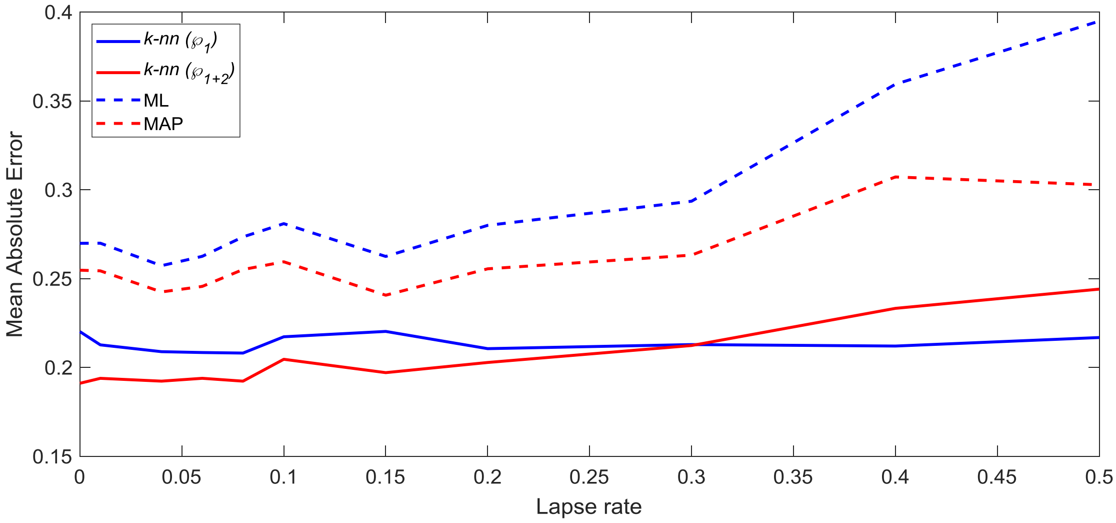

Fig 4. MAE of extracted *w* by *k*-nn and fitting *in* the presence of lapse. Each point represents 10000 independent agents that perform the task by the DS-Eλ-DS and random parameters. After making a decision, it toggled by the probability of lapse ratio. The fitted model is chosen based on AIC. The *k*-nn estimation is applied by all the different feature spaces mentioned before. The confidence interval is very close to the results.

Based on Fig 4, it's clear that k-nn methods are more resistant to lapse than traditional fitting methods, especially when fitting-based features are taken out of the feature space.

4 Experimental Data Analyses

This section validates the proposed method using actual experimental data. To validate the proposed method, data from two independent research were chosen. The comparison of results based on "combination weight from the proposed approach" ($\hat{w}_{k\text{-nn}}$) to results based on "combination weight from traditional methods" ($\hat{w}_{\mathrm{ML}}$ or $\hat{w}_{\mathrm{MAP}}$) demonstrates the superiority of the proposed method.

4.1 Analysis of relationship between Learning style and Gaze direction

Using the Daw task, Konovalov and Krajbich have already looked into the correlation between gaze information and combination weight (w) (Konovalov & Krajbich, 2016). They use the Daw task in two ways, and we use the first one to make sure our models are the same. ML has used the IS-λE-SS version of the model to get the w value (see Table 1). In this study, we used the k-nn estimation with $\wp_{1+2}$ feature space to extract the w from their data. The number of trials in the Konovalov study (Konovalov & Krajbich, 2016) has been set to 150, so we make a different database by setting number of trials in simulations to 150.

Konovalov and Krajbich divided subjects into two groups based on the median of $\hat{w}_{ML}$ (0.3) to study the differences between MB and MF behavior. When the $\hat{w}_{k\text{-nn}}$ instead of the $\hat{w}_{\mathrm{ML}}$ was utilized, several subjects' groupings were altered. We first focus on the behavioral differences in sense of stay probability in different groups and those subjects that the traditional method and proposed method have conflict in the grouping. The analysis indicates that the traditional method divides the subject better than the traditional one in the sense of stay probability (See supplementary 4).

We validate all of the studies presented in the first part of the paper by Konovalov and Krajbich using k-nn group labels (Konovalov & Krajbich, 2016). While the major analytical results remained unchanged, there were an outstanding relationship. We examine the correlation between $\hat{w}_{k\text{-nn}}$ and all behavioral data of subjects. There is no correlation between $\hat{w}_{\mathrm{ML}}$ and available meaningful behaviorally indices that not mentioned in the paper. But using the proposed method, we observe that the mean dwell time in middle gazes was strongly correlated with $\hat{w}_{k\text{-nn}}$ (correlation coefficient = 0.5, p-value = 0.001). In contrast, the $\hat{w}_{\mathrm{ML}}$ and the mean dwell time of middle gazes were not correlated (correlation coefficient =0.08, p-value =0.603). Fig 5 depicts this amazing association.

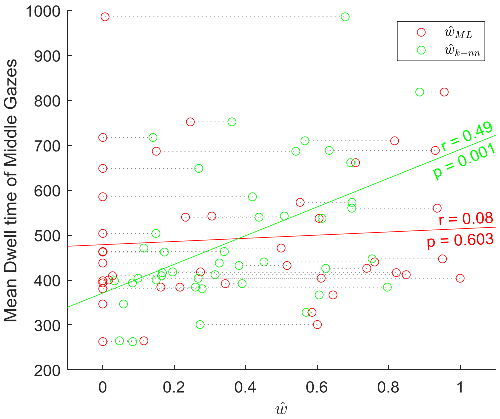

Fig 5 The correlation of the mean dwell time in middle gazes and traditionally fitted *w* (red) and estimated w by the proposed method (green). The corresponding correlation coefficients and p-values are reported in the graph.

The proposed method shows some information from the data that would have been missed if traditional fitting methods were used. This information did not change the results of (Konovalov & Krajbich, 2016) study.

4.2 Analysis of the Relationship between Learning Style and Symptom Dimension

Gillan et al. used the Daw task to examine the relationship between learning style and compulsive behaviors (Gillan et al., 2016). While they utilize the Daw task without modification, their analytical model differs. Their computational model is a modified version of the reparameterization model presented by Otto et al (Otto et al., 2013). They demonstrated strong correlations between certain psychiatric diseases and the subject's preference for MB style. However, these correlations for the combination weight are absent due to imprecise estimation in conventional fitting methods (see below). We believe a more accurate estimation strategy can revive these relationships in the Daw et. al. model.

To have a fair comparison, we do the same analysis as Gillan et al. but the model version assumed DS-λE-1S instead of a modified reparameterization model. We use traditional fitting methods and the proposed method to extract $\hat{w}$. Table 6 report the correlation between subject aspects and reports $\hat{w}$ also the Table 6 report the regression analysis between the self-report questionnaire's total scores and $\hat{w}$.

As a control for regression analysis, Gillan et al. used age, IQ, and gender, which have been previously reported to covary with goal-directed behavior (Eppinger et al., 2013; Gillan et al., 2016; Schad et al., 2014). In line with the Gillan et al. study, the extracted combination weight by *k*-nn methods has significant relationships with age, IQ, and Gender, but there is only a relationship between age and $\hat{w}_{\mathrm{ML}}$ (see Table 6 for more details). The traditional fitting method extracts the ŵ that is not consistent with other analyses like the Eppinger et al. research or the Schad et al. study(Eppinger et al., 2013; Schad et al., 2014).

Table 5. Correlation between Age, Gender, and IQ z-score and combination weight (correlation coefficient (p-value))

(yellow box significant by level 0.05 and green box significant by level 0.01)

	k-nn ($\wp_{1+2}$)	*k*-nn ($\wp_1$)	ML	MAP
Age	-0.162 (9.7e-10)	-0.058 (0.029)	-0.059 (0.028)	0.010 (0.704)
Gender	0.115 (1.6e-05)	0.084 (0.002)	0.029 (0.282)	0.020 (0.446)
IQ	0.237 (2.0e-19)	0.163 (7.1e-10)	0.036 (0.178)	-0.017 (0.529)

Based on one trial back regression analysis in the Gillan et al study, there was a significant inverse association between goal-directed behavior and scores on the eating disorder, Impulsivity, OCD, and alcohol addiction questionnaire (see Table 6 for more details). The *k*-nn methods replicated some of this association, but the traditional fitting methods did not. The *k*-nn ($\wp_{1+2}$) replicates the association between goal-directed behavior and score of OCD and alcohol addiction. Also, the *k*-nn ($\wp_1$) method replicates the association between goal-directed behavior and the score of impulsivity, and alcohol addiction (see Table 6 for more details). On the other hand, the MAP method replicates an association between goal-directed parameters and apathy score which is not in line with other studies and regression analyses.

Gillan et al. introduced three factors for more analysis, and we also analyzed the correlation between these factors and extracted $\hat{w}$ by different methods. The regression analysis shows a significant association between factor 2 or 'Compulsive Behavior and Intrusive Thought' and goal-directed behavior (β =-0.046, SE=0.01, p<0.001). The proposed method also replicates this relationship, but the traditional fitting methods missed it. 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.

The proposed method could replicate some relationships between goal-directed behavior and some psychiatric disorders, but 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 a different model. So we can claim that this estimation method is more reliable than traditional methods in finding clinically relevant relationships.

Table 6. Regression Analysis of self-report questionnaire total z-score and combination weight

The first column is One-Trial-Back regression (symptom_score_z ~ Reward * Transition * Stay + Reward * Transition * (IQ_z + Age_z + Gender) + (Reward * Transition + 1| Subject))

Each raw of last 4 columns replicate the regression analysis of symptom_score_z ~ 1+Age_z +IQ_z +Gender+ $\hat{w}$.

(yellow box significant by level 0.05 and green box significant by level 0.01)

Clinical Scores	β (p-value) [residual]				
	One-Trial Back Regression (replicated of(Gillan et al., 2016))	k-nn (℘1+2)	k-nn (℘1)	ML	MAP
Eating Disorders	-0.041 (<.001) [.042]	-0.036 (0.163) [0.082]	-0.010 (0.576) [0.032]	-0.017 (0.844) [0.005]	0.001 (0.215) [0.002]
Impulsivity	-0.039 (.002) [.028]	-0.035 (0.180) [0.082]	-0.011 (0.002) [0.039]	-0.019 (0.125) [0.006]	0.001 (0.139) [0.002]
OCD	-0.03 (.018) [.050]	-0.036 (0.038) [0.083]	-0.010 (0.155) [0.033]	-0.017 (0.648) [0.005]	0.002 (0.213) [0.002]
Alcohol Addiction	-0.03 (.029) [.052]	-0.036 (0.026) [0.084]	-0.011 (0.049) [0.034]	-0.019 (0.084) [0.007]	0.004 (0.513) [0.001]
Schizotypy	-0.02 (.101) [.028]	-0.035 (0.516) [0.081]	-0.010 (0.216) [0.033]	-0.017 (0.782) [0.005]	0.001 (0.090) [0.003]
Depression	-0.01 (.351) [.031]	-0.034 (0.608) [0.081]	-0.009 (0.724) [0.032]	-0.017 (0.783) [0.005]	0.001 (0.197) [0.002]
Trait Anxiety	-0.01 (.552) [.038]	-0.034 (0.932) [0.080]	-0.009 (0.899) [0.032]	-0.017 (0.903) [0.005]	0.001 (0.260) [0.002]
Apathy	-0.00 (.897) [.015]	-0.033 (0.260) [0.081]	-0.009 (0.845) [0.032]	-0.016 (0.112) [0.007]	0.002 (0.007) [0.006]
Social Anxiety	0.01 (.503) [.028]	-0.034 (0.666) [0.081]	-0.009 (0.955) [0.032]	-0.017 (0.931) [0.005]	0.002 (0.092) [0.003]
Factors					
'Anxious-Depression'	-0.02 (.967) [.018]	-0.033 (0.528) [0.081]	-0.009 (0.708) [0.032]	-0.017 (0.886) [0.005]	0.001 (0.062) [0.003]
'Compulsive Behavior and Intrusive Thought'	-0.061 (<.001) [.088]	-0.039 (0.005) [0.086]	-0.013 (0.029) [0.035]	-0.019 (0.523) [0.005]	0.001 (0.422) [0.001]
'Social Withdrawal'	0.03 (.282) [.036]	-0.034 (0.960) [0.080]	-0.009 (0.559) [0.032]	-0.017 (0.708) [0.005]	0.002 (0.049) [0.004]

5 Discussion and Conclusion

The MB and MF learning balance extraction is necessary for the transition of reinforcement learning modeling to mathematical psychology. The Daw two-step task was designed to disassociate MB and MF learning styles and was used widely. We study the precision of extracting the subject's preference towards MB style using this task. We used nine nested versions of the model. 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.

Our analysis specified that the complex model over-fit to the observation and simple models with erroneous assumptions lead to higher errors (see supplementary 2 for details). Moreover, when prior knowledge was not assumed for the fitted parameters, sometimes the fitted values stick to the extremes of the parameter range. Our analysis shows that the agent parameter also affects error. MB and MF styles have similar behavior when the learning rate or inverse temperature is low. In these conditions, the estimation error increase (see supplementary 2 for details). Such problems in model fitting make the fitted parameters unreliable (Eckstein et al., 2022).

Besides the traditional model fitting, some statistical indices were extracted and used for investigating the cognitive studies from the behavioral data. 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 learn the parameter estimation instead of model fitting. We use 20 features (including fitting-based features) to generate the *k*-nn dataset and then, we extract two different feature-spaces by elimination of fitting-based features. Eliminating the fitted-based features reduces both computational load and noise effect. The best performance was reached by *k*-nn. Both bias and variance of error were proven to be reduced by *k*-nn learning compared to traditional model fitting. The analysis also specifies that the k nn method is more stable in the presence of lapse, especially by excluding all fitting-based features. when we use fitting-based features, we involved the model fittings problems such as low sample size, selecting a good model, optimization method, and objective function, so if we have no information about the model or fitting, it is better to ignore the 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, making the proposed method superior performance over just model-fitting approaches. MAP estimation is better than ML in extreme values because using a prior, *k*-nn method works better than MAP. The mentioned improvements will enhance the applicability of the Daw task for computational psychiatry purposes.

It was indicated that using the proposed method can help to find a significant correlation between w and mean dwell time which is missing in the traditional method. 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, 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 and GLMM in Gillan et al. study (Gillan et al., 2016). Because adding some noise to one variable can destroy the correlation coefficient between that variable and other measures, some correlation coefficient has lost their significance due to noisy estimation of combination weight, and the proposed method was more successful due to the reduction of this noise. It is worth noting that though the proposed method successfully extracted most relationships of the Gillan et al. study, some relationships were missing even by *k*-nn. 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 minimizes an objective function to extract the behavior under different assumptions. The ML 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 a specific parameter to not guarantee maximum likelihood while another parameter exists that satisfies the maximized likelihood criterion. Although ML can theoretically obtain the Cramer-Rao Lower Band, the above statement is the cause that learning reaches better estimations rather than ML. The proposed method can be considered a maximum likelihood estimation using simulation-based estimation. Such a method uses trial-by-trial observations of the behavior and global observation such as stay probabilities in random variable space and tries to maximize the likelihood of observing all the mentioned behaviors together. ML and *k*-nn methods may converge to the same estimation error for large sample sizes. However, for limited sample sizes, *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 model-based and model-free combination weight estimation. 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 diagnose psychiatric disorders.

Acknowledgment

We would like to express our gratitude to Arkady Konovalov and Ian Krajbichv for sharing the gaze task data with us and Gillan et al.

Availability of data and materials

The human behavior data Gaze Direction (Konovalov & Krajbich, 2016) and Symptom Dimension (Gillan et al., 2016) are available from the authors of that papers. The codes and datasets generated during the current study are accessible via Dataverse repository: https://doi.org/10.7910/DVN/PSEFZF.

References

Ahn, W. Y., & Busemeyer, J. R. (2016). Challenges and promises for translating computational tools into clinical practice. *Current Opinion in Behavioral Sciences*, *11*, 1–7. https://doi.org/10.1016/j.cobeha.2016.02.001

Collins, A. G. E., Albrecht, M. A., Waltz, J. A., Gold, J. M., & Frank, M. J. (2017). Interactions among working memory, reinforcement learning, and effort in value-based choice: A new paradigm and selective deficits in schizophrenia. *Biological Psychiatry*, *82*(6), 431–439. https://doi.org/10.1016/j.biopsych.2017.05.017

Culbreth, A. J., Westbrook, A., Daw, N. D., Botvinick, M., & Barch, D. M. (2016). Reduced model-based decision-making in schizophrenia. *Journal of Abnormal Psychology*, *125*(6), 777–787. https://doi.org/10.1037/abn0000164

Daw, N. D. (2015). Of goals and habits. *Proceedings of the National Academy of Sciences*, *112*(45), 13749–13750. https://doi.org/10.1073/pnas.1518488112

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711. https://doi.org/10.1038/nn1560

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312–325. https://doi.org/10.1016/j.neuron.2013.09.007

Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D., & Daw, N. D. (2015). Model-based choices involve prospective neural activity. *Nature Neuroscience 2015 18:5*, *18*(5), 767–772. https://doi.org/10.1038/nn.3981

Eckstein, M. K., Master, S. L., Xia, L., Dahl, R. E., Wilbrecht, L., & Collins, A. G. (2022). The interpretation of computational model parameters depends on the context. *ELife*, *11*. https://doi.org/10.7554/elife.75474

Eppinger, B., Walter, M., Heekeren, H. R., & Li, S. C. (2013). Of goals and habits: Age-related and individual differences in goal-directed decision-making. *Frontiers in Neuroscience*, *7*(7 DEC), 1–14. https://doi.org/10.3389/fnins.2013.00253

Everitt, B. J., & Robbins, T. W. (2005). Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. *Nature Neuroscience*, *8*(11), 1481–1489. https://doi.org/10.1038/nn1579

Feher da Silva, C., & Hare, T. A. (2020). Humans primarily use model-based inference in the two-stage task. *Nature Human Behaviour*, *4*(10), 1053–1066. https://doi.org/10.1038/s41562-020-0905-y

Foerde, K. (2018). What are habits and do they depend on the striatum? A view from the study of neuropsychological populations. *Current Opinion in Behavioral Sciences*, *20*, 17–24. https://doi.org/10.1016/J.COBEHA.2017.08.011

Gijsen, S., Grundei, M., & Blankenburg, F. (2022). Active inference and the two ‐ step task. *Scientific Reports*, *0123456789*, 1–15. https://doi.org/10.1038/s41598-022-21766-4

Gillan, C. M., & Daw, N. D. (2016). Taking Psychiatry Research Online Claire. *Neuron*, *91*(1), 19–23. https://doi.org/10.1016/j.neuron.2016.06.002

Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A., & Daw, N. D. (2016). Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *ELife*, *5*, 1–24. https://doi.org/10.7554/eLife.11305

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. https://doi.org/10.3758/s13415-015-0347-6

Gillan, C. M., Papmeyer, M., Morein-Zamir, S., Sahakian, B. J., Fineberg, N. A., Robbins, T. W., & De Wit, S. (2011). Disruption in the balance between goal-directed behavior and habit learning in obsessive-compulsive disorder. *American Journal of Psychiatry*, *168*(7), 718–726. https://doi.org/10.1176/appi.ajp.2011.10071062

Gillan, C. M., & Robbins, T. W. (2014). Goal-directed learning and obsessive-compulsive disorder. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655), 20130475. https://doi.org/10.1098/rstb.2013.0475

Keramati, M., Smittenaar, P., Dolan, R. J., & Dayan, P. (2016). Adaptive integration of habits into depth-limited planning defines a habitual-goal–directed spectrum. *Proceedings of the National Academy of Sciences*, *113*(45), 12868–12873. https://doi.org/10.1073/pnas.1609094113

Konovalov, A., & Krajbich, I. (2016). Gaze data reveal distinct choice processes underlying model-based and model-free reinforcement learning. *Nature Communications*, *7*(C), 12438. https://doi.org/10.1038/ncomms12438

Kool, W., Cushman, F. A., & Gershman, S. J. (2016). When does model-based control pay off? *PLoS Computational Biology*, *12*(8), 1–34. https://doi.org/10.1371/journal.pcbi.1005090

Kroemer, N. B., Lee, Y., Pooseh, S., Eppinger, B., Goschke, T., & Smolka, M. N. (2019). L-DOPA reduces model-free control of behavior by attenuating the transfer of value to action. *NeuroImage*, *186*, 113–125. https://doi.org/10.1016/J.NEUROIMAGE.2018.10.075

Li, Z., Liu, G., & Li, Q. (2017). Nonparametric Knn estimation with monotone constraints. *Econometric Reviews*, 1–19.

Lucantonio, F., Caprioli, D., & Schoenbaum, G. (2014). Transition from 'model-based' to 'model-free' behavioral control in addiction: involvement of the orbitofrontal cortex and dorsolateral striatum. *Neuropharmacology*, *23*(1), 1–7. https://doi.org/10.1016/j.neuropharm.2013.05.033.

Miller, K. J., Botvinick, M., Brody, C. D., Miller, K. J., Brody, C. D., & Botvinick, M. M. (2016). Identifying model-based and model-free patterns in behavior on multi-step tasks. *BioRxiv*, 096339. https://doi.org/10.1101/096339

Miller, K. J., Botvinick, M. M., & Brody, C. D. (2022). Value representations in the rodent orbitofrontal cortex drive learning, not choice. *ELife*, *11*, 1–27. https://doi.org/10.7554/eLife.64575

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2013). Computational psychiatry. *Trends in Cognitive Sciences*, *16*(1), 72–80. https://doi.org/10.1016/j.tics.2011.11.018

Morris, L. S., Baek, K., & Voon, V. (2017). Distinct cortico-striatal connections with subthalamic nucleus underlie facets of compulsivity. *Cortex*, *88*, 143–150. https://doi.org/10.1016/J.CORTEX.2016.12.018

Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A., & Daw, N. D. (2013). Working-memory capacity protects model-based learning from stress. *Proceedings of the National Academy of Sciences*, *110*(52), 20941–20946. https://doi.org/10.1073/pnas.1312011110

Schad, D. J., Jünger, E., Sebold, M., Garbusow, M., Bernhardt, N., Javadi, A. H., Zimmermann, U. S., Smolka, M. N., Heinz, A., Rapp, M. A., & Huys, Q. J. M. (2014). Processing speed enhances model-based over model-free reinforcement learning in the presence of high working memory functioning. *Frontiers in Psychology*, *5*(DEC), 1–10. https://doi.org/10.3389/fpsyg.2014.01450

Shahar, N., Hauser, T. U., Moutoussis, M., Moran, R., Keramati, M., Dolan, R. J., & Dolan, R. J. (2019). Improving the reliability of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-diffusion modeling. *PLOS Computational Biology*, *15*(2), e1006803. https://doi.org/10.1371/journal.pcbi.1006803

Smittenaar, P., FitzGerald, T. H. B., Romei, V., Wright, N. D., & Dolan, R. J. (2013). Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans. *Neuron*, *80*(4), 914–919. https://doi.org/10.1016/j.neuron.2013.08.009

Toyama, A., Katahira, K., & Ohira, H. (2017). A simple computational algorithm of model-based choice preference. *Cognitive, Affective and Behavioral Neuroscience*, *17*(4), 764–783. https://doi.org/10.3758/s13415-017-0511-2

Toyama, A., Katahira, K., & Ohira, H. (2019). Biases in estimating the balance between model-free and model-based learning systems due to model misspecification. *Journal of Mathematical Psychology*, *91*, 88–102. https://doi.org/10.1016/J.JMP.2019.03.007

Voon, V., Derbyshire, K., Rück, C., Irvine, M. A., Worbe, Y., Enander, J., Schreiber, L. R. N., Gillan, C. M., Fineberg, N. A., Sahakian, B. J., Robbins, T. W., Harrison, N. A., Wood, J., Daw, N. D., Dayan, P., Grant, J. E., & Bullmore, E. T. (2015). Disorders of compulsivity: a common bias towards learning habits. *Molecular Psychiatry*, *20*(3), 345–352. https://doi.org/10.1038/mp.2014.44

Wanjerkhede, S. M., Bapi, R. S., & Mytri, V. D. (2014). Reinforcement learning and dopamine in the striatum: A modeling perspective. *Neurocomputing*, *138*, 27–40. https://doi.org/10.1016/j.neucom.2013.02.061

Ward, M. D., Carolina, N., & Ahlquist, J. S. (2012). *Maximum likelihood for social sciences strategies for analysis* (Issue April).

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*(8), 1293–1313. https://doi.org/10.3758/BF03194544

Wit, S. de, Barker, R. A., Dickinson, A. D., & Cools, R. (2011). Habitual versus goal-directed action control in parkinson disease. *Journal of Cognitive Neuroscience*, *23*(5), 1218–1229. https://doi.org/10.1162/jocn.2010.21514