# Extended Bayesian inference incorporating symmetry bias

Shuji Shinohara[1*], Nobuhito Manome[1,2], Kouta Suzuki[1,2], Ung-il Chung[1], Tatsuji Takahashi[1,3], Pegio-Yukio Gunji[4], Yoshihiro Nakajima[5], Shunji Mitsuyoshi[1]

[1]Department of Bioengineering, Graduate School of Engineering, The University of Tokyo, Tokyo, Japan

[2] Department of Research and Development, SoftBank Robotics Group Corp, Tokyo, Japan

[3]School of Science and Engineering, Tokyo Denki University, Saitama, Japan

[4]Department of Intermedia Art and Science, School of Fundamental Science and Technology, Waseda University, Tokyo, Japan

[5]Graduate School of Economics, Osaka City University, Osaka, Japan

* Corresponding author

E-mail: shinohara@bioeng.t.u-tokyo.ac.jp (SS)

## Abstract

In this study, we start by proposing a causal induction model that incorporates symmetry bias. This model is important in two aspects. First, it can reproduce causal induction of human judgment with higher accuracy than conventional models. Second, it allows us to estimate the level of symmetry bias of subjects from experimental data. We further propose an inference method that incorporates the aforementioned causal induction model into Bayesian inference. In this method, the component of Bayesian inference, which updates the degree of confidence for each hypothesis, and the component of inverse Bayesian inference that modifies the model of the hypothesis coexist. Our study demonstrates that inverse Bayesian inference enables us to deal flexibly with unstable situations where the object of inference changes from time to time.



## Author summary

We acquire knowledge through learning and make various inferences based on such knowledge and observational data (evidence). If the evidence is insufficient, then the certainty of the conclusion will decline. Moreover, even if the evidence is sufficient, the conclusion may be wrong if the knowledge is incomplete in the first place. In order to model such inference based on incomplete knowledge, we proposed an inference system that performs learning and inference simultaneously and seamlessly. Prepare two coins A and B with different probabilities of landing heads, and repeat the coin toss using either of them. However, the coin that is being tossed is also replaced repeatedly. The system observes only the result of coin toss each time, and estimates the probability of landing heads of coin tossed at the moment. In this task, it is necessary not only to estimate the probabilities of the landing heads of coin A and B, but also to estimate which coin is being used at the moment. In this paper, we show that the proposed system handles such tasks very efficiently by simultaneously performing inference and learning.


## Introduction

As a cognitive bias observed in humans, the disposition to infer from 'if P then Q' to 'if Q then P' or to 'if not P, then not Q' is well documented [1, 2, 3, 4, 5, 6, 7, 8]. The former is termed symmetry bias [9] and the latter is termed mutual exclusivity bias [4].

Consider a simple example. We tend to infer from 'if you clean the room, then I will take you out' to 'I will take you out if and only if you clean the room' or 'if you don't clean the room, then I will not take you out'. Although these inferences are invalid according to classical logic, various people are inclined to make them regardless of age.

In contrast, among non-human animals, the symmetry bias has been reported in behaviour of only some California sea lions [10] and chimpanzees [11]. Although the symmetry bias produces wrong inferences from the classical logic point of view, humans do show some positive features apparently stemming from the symmetry bias. For instance, once you are able to respond to the question 'What is this?' with 'This is an apple' through learning, you will also be able to identify the correct object when asked 'Which one is an apple?'. In other words, we automatically infer from 'This is an apple' to 'An apple is this' without any instruction. The symmetry bias has been studied in relation to stimulus equivalence in the field of comparative psychology [1, 2]. On the other hand, the mutual exclusivity bias has been studied primarily in the field of developmental psychology in the context of young children's language acquisition [4]. Thus, although the symmetry and mutual exclusivity biases have been studied in different fields of psychology, since the contrapositive of 'if Q then P' is 'if not P then not Q', and since they are equivalent according to classical logic, the same implications can be associated with both biases. An example of such a shared implication is that both biases may be caused by the same neuroscientific factor.

Concurrently, in the field of cognitive psychology, experiments on causal induction were carried out, seeking to identify how humans evaluate the strength of causal relations between two events. In a regular conditional statement of the form 'if $p$ then $q$', the degree of confidence is considered to be proportional to the conditional probability $P(q \mid p)$ which is the probability of occurrence of $q$ following the occurrence of $p$ [12]. Further, in the case of causal relation, it has been experimentally demonstrated that humans have a

strong sense of causal relation when $P(p \mid q)$ is high, as well as when $P(q \mid p)$ is high, where $P(p \mid q)$ is a conditional probability of the antecedent occurrence of $p$, given the occurrence of $q$ [13].

Consider a simple causal induction model that infers the strength of a causal relation from the cause candidate of event $C$ to the effect event $E$ from four pieces of co-occurring information concerning $C$ and $E$: the joint presence of $C$ and $E$, absence of $E$ given $C$, presence of the $E$ given no $C$, and the joint absence of $C$ and $E$. The most representative model of causal induction is the $\Delta P$ model [14]. It takes the difference between the conditional probability $P(E \mid C)$ of occurrence of $E$ given the occurrence of $C$ and the conditional probability $P(E \mid \neg C)$ of occurrence of $E$ given non-occurrence of $C$ (denoted by $\neg C$) as an index for causal strength, that is, $\Delta P = P(E \mid C) - P(E \mid \neg C)$.

Hattori and Oaksford proposed the dual-factor heuristic (*DFH*) model [13]. This model is based on the geometric mean of $P(E \mid C)$, which stands for the predictability of the effect from the cause, and its inverse $P(C \mid E)$, that is, $DFH = \sqrt{P(E \mid C)P(C \mid E)}$.

Both $\Delta P$ and $DFH$ models contain $P(E \mid C)$. In other words, given the occurrence of $C$, if the probability of occurrence of $E$ following $C$ is high, the chance of $C$ to be the cause of $E$ increases. Intuitively speaking, the strength of the causal relation does not seem to be solely determined by $P(E \mid C)$. The second item in the $\Delta P$ model, $-P(E \mid \neg C)$, shows that even if the probability of occurrence of $E$ is high given the occurrence of $C$, if the probability of occurrence of $E$ is still high in the absence of the occurrence of $C$, that is, if the probability of occurrence of $E$ is high irrespectively of the occurrence of $C$, then the chance of $C$ being the cause of $E$ decreases.

Whereas for the $DFH$ model, if the probability $P(C \mid E)$, which is the probability of the antecedent occurrence of $C$ given the occurrence of $E$, is high, the chance of $C$ being the cause of $E$ increases. This can be understood as a probabilistic expression of the belief that where there is no cause, there is no effect.

We can also consider the $\Delta P$ model and the $DFH$ model in terms of biases. For the sake of simplicity, $\Delta P$ and $DFH$ are expressed as $\Delta P(E|C)$ and $DFH(E|C)$ respectively. Here, if we assign $\neg C$

and $\neg E$ to $C$ and $E$ respectively in $\Delta P(E|C) = P(E \mid C) - P(E \mid \neg C)$, we can obtain mutual exclusivity as $\Delta P(\neg E \mid \neg C) = P(\neg E \mid \neg C) - P(\neg E \mid \neg(\neg C)) = (1 - P(E \mid \neg C)) - (1 - P(E \mid C)) = P(E \mid C) - P(E \mid \neg C) = \Delta P(E \mid C)$. If, on the other hand, $C$ in $DFH(E|C) = \sqrt{P(E \mid C)P(C \mid E)}$ is replaced by $E$, we can get $DFH(C \mid E) = \sqrt{P(C \mid E)P(E \mid C)} = \sqrt{P(E \mid C)P(C \mid E)} = DFH(E \mid C)$ and the symmetry obtains.

Aside from the $\Delta P$ and $DFH$ models, Takahashi and colleagues [15] proposed the *pARIs* (proportion of assumed-to-be rare instances) as yet another model that has an unusually high affinity with the human causal induction judgment, $pARIs(E|C) = P(C,E)/(P(C,E) + P(C,\neg E) + P(\neg C,E))$, where $P(x,y)$ represents joint probability of $x$ and $y$. If $C$ in $pARIs(E|C)$ is replaced by $E$, we get $pARIs(C|E) = P(E,C)/(P(E,C) + P(E,\neg C) + P(\neg E,C)) = pARIs(E|C)$ and the symmetry obtains.

To view the relation between two events as a causal relation can therefore be understood as having both symmetry and mutual exclusivity biases.

Bayesian inference is based on the notion of conditional probability. Bayesian inference speculates the hidden cause behind an observation results from retrospectively applying statistical inferences. The relation between Bayesian inference and brain function has been attracting attention in recent years in the field of neuroscience [16, 17].

In Bayesian inference, the degree of confidence in a hypothesis is updated based on a model of predefined hypotheses and current observational data. In other words, Bayesian inference is a process of narrowing down hypotheses to one which best explains observational data. Changing the model of each hypothesis or adding new ones in the course of performing Bayesian inference is not allowed. In addition, Bayesian inference itself does not deal with alterations in the inference target during the inference or with its replacement. Therefore, such inference substantially needs to assume the identity of the target.

Note, however, that requirements of the invariability of the hypothetical model and the identity of the inference target stem from the theoretical framework, and they are not always met in actuality. For instance, if the object is unknown, it would be impossible to infer what it is without adding a new hypothetical

model. Moreover, it is likely that, under unsteady circumstances, the inference target undergoes alteration from time to time or is replaced by some other object.

In order to predetermine whether the object is replaced by another, one must first infer its identity. A correct inference depends on as much observational data as possible. However, in order to properly use accumulated observational data, it must be ensured that these data derive from the same object. In other words, to determine whether the object has been replaced or not, the object must be hypothesized not to have been replaced in the first place. In this sort of situation, it is necessary to infer what the object is while at the same time evaluating the legitimacy of the inference itself. How, then, could we model the inference under the situation described above?

Arecchi [18] proposed the concept of the inverse Bayesian inference where the hypothetical model, which is fixed in the traditional Bayesian inference, is modified according to circumstances. Gunji et al. [19, 20] and Horry et al. [21] formulated the inverse Bayesian inference and demonstrated that animal herding and human decision-making can be satisfactorily modelled by combining Bayesian inference and inverse Bayesian inference. This framework can be said to seamlessly perform Bayesian inference, by picking up the optimal hypothesis from the predefined set of hypotheses, and simultaneously apply inverse Bayesian inference (learning), which creates a new hypothesis according to observational data. Although the inverse Bayesian inference was formulated by Gunji and others [19, 20], it is not necessarily linked with causal inference and symmetry bias.

We propose a causal induction model that primarily incorporates symmetry bias. First, we propose an extended model of degree of confidence that extends conditional probability by parametrising the mixed rate of $P(q \mid p)$ and $P(p \mid q)$, i.e., the strength of symmetry bias. Second, we propose a realistic human inference model that incorporates the extended model into Bayesian inference, and we show that it necessarily involves inverse Bayesian inference. Specifically, we propose a framework of extended Bayesian inference which allows seamless and simultaneous learning and inference by replacing the conditional probability schema in Bayesian inference with the extended model of degree of confidence. Third, we explain

a conducted simulation, derived from the problem of inference, of the probability of getting heads in the course of repetitive coin toss and how it helps verify the legitimacy of the Extended Bayesian Inference.

# Results

## Proposal of extended confidence model

We seek to establish an extended model of degree of confidence as the generalised weighted average of $P(q \mid p)$ and its inverse $P(p \mid q)$ using parameters $\alpha$ and $m$.

$$C(q \mid p) = [(1-\alpha)P(q \mid p)^m + \alpha P(p \mid q)^m]^{1/m} \qquad (1)$$

Hereinafter, $C(q \mid p)$ will be termed the Extended Confidence Model. Here $\alpha$ takes values in the range $0.0 \leq \alpha \leq 1.0$ and denotes a weighted value of $P(q \mid p)$ and $P(p \mid q)$, and $m$ takes values in the range $-\infty \leq m \leq \infty$ and denotes the manner of taking the mean. For example, suppose $\alpha = 0.5$ and $m = 1.0$, then $C(q \mid p) = 0.5P(q \mid p) + 0.5P(p \mid q)$ representing the arithmetic mean. Supposing $m = 0.0$, the formula (1) is undefinable. If, however, we represent the mean value in the limit of $m \to 0.0$, we get $C(q \mid p) = P(q \mid p)^{1-\alpha}P(p \mid q)^{\alpha}$, where if $\alpha = 0.5$, the geometric mean $C(q \mid p) = \sqrt{P(q \mid p)P(p \mid q)}$ it coincides with the *DFH* model. If $\alpha = 0.5, m = -1.0$, the formula represents the harmonic mean of $P(q \mid p)$ and $P(p \mid q)$ and we get

$$C(q \mid p) = \frac{2P(q \mid p)P(p \mid q)}{P(q \mid p) + P(p \mid q)} = \frac{2P(p,q)}{2P(p,q) + P(p,\neg q) + P(\neg p,q)} = \frac{2pARIs}{1 + pARIs} \qquad (2)$$

and $C(q \mid p)$ can be expressed as a harmonic mean of 1 and $pARIs$. In other words, $pARIs$ and $C(q \mid p)$ are related by monotonically increasing functions that are in one-to-one correspondence. According to this, $pARIs$ can be seen as a disguised form of $C(q \mid p)$. Here, the parameter $\alpha$ can be regarded as a parameter

that controls that strength of the symmetry bias. When $\alpha = 0$, $C(q \mid p) = P(q \mid p)$ obtains irrespectively of the value of $m$, and $C(q \mid p)$ expresses a normal conditional probability without the symmetry bias.

Thus, the proposed model can be described as an extended model which accommodates the normal conditional probability $P$, $DFH$ and $pARIs$ as inner special cases.

## Evaluation of descriptive validity of extended confidence model

In order to evaluate the descriptive validity of various models including the *DFH* model, Hattori and Oaksford [13] performed meta-analysis using data from eight types of causal induction experiments. To test the descriptive performance of the extended confidence model, we also performed the meta-analysis using the same datasets as Hattori and Oaksford [13].

Generally, in a simple causal induction experiment, participants are given four types of co-occurrence information concerning the cause $C$ and the effect $E$ (Table 1). Then, they are asked to assess subjectively the strength of the causal relation between $C$ and $E$ using a number from 0 to 100. To measure each model's fit to the data, we calculated the determination coefficient $R^2$ from the pair of participants' mean ratings of causal strength and the estimated value of each model computed from the same co-occurrence information given to the participants in the experiments.

**Table 1. The 2 × 2 contingency table for elemental causal induction.**

| | effect ($E$) | no effect ($\neg E$) | marginal frequency |
|---|---|---|---|
| cause ($C$) | $N(C,E)$ | $N(C,\neg E)$ | $N(C)$ |
| no cause ($\neg C$) | $N(\neg C,E)$ | $N(\neg C,\neg E)$ | $N(\neg C)$ |
| marginal frequency | $N(E)$ | $N(\neg E)$ | |

N(x) denotes the frequency of occurrence of x, and N(x,y) denotes the frequency of co-occurrence of x and y.

The experiment data used in the meta-analysis consists of the experiment I from [22], experiments I and III from [23], experiments I and II from [13], experiments I and III from [24], and experiment II and VI from [25]. The experiment data above will be abbreviated as AS95, BCC03.1, BCC03.3, HO07.1,

HO07.2, LS00, W03.2, and W03.6 respectively. See the methods section below for the content of each experiment.

Values in parameters $\alpha$ and $m$ in formula (1) were shifted each with 0.05 increment in the interval $[0.0,1.0]$ and $[-2.0,2.0]$ and the determination coefficient $R^2$ between assessment by participants and the estimated value by the proposed model was calculated for each pair of parameters. Table 2 shows the pair of parameters $\alpha$ and $m$ at which $R^2$ becomes maximum in each experiment. As seen in Table 2, the determination coefficients were greater than 0.9 for all experiments. $\alpha$ was around 0.5 (0.25-0.6) and did not reach 0.0, which stands for the normal conditional probability *P*. This suggests that symmetry bias was deeply involved in causal induction. Moreover, $m$ took a negative value in all experiments. This suggest people show strong awareness of causal relations only if both $P(E \mid C)$ and its inverse $P(C \mid E)$ are large.

**Table 2. Performance evaluation of the extended confidence model based on the meta-analytic data from Hattori and Oaksford (13).**

| | AS95 | BCC03.1 | BCC03.3 | HO07.1 | HO07.2 | LS00 | W03.2 | W03.6 |
|---|---|---|---|---|---|---|---|---|
| $R^2$ | 0.97 | 0.98 | 0.99 | 0.99 | 0.99 | 0.91 | 0.97 | 1.00 |
| $\alpha$ | 0.3 | 0.55 | 0.35 | 0.6 | 0.3 | 0.55 | 0.35 | 0.25 |
| $m$ | -0.65 | -0.45 | -0.25 | -2.0 | -2.0 | -1.4 | -1.6 | -2.0 |
| $N$ | 80 | 13 | 6 | 12 | 9 | 11 | 8 | 4 |

$R^2$ is a determination coefficient, $\alpha$ and $m$ are model parameters, and *N* is a number of combinations of stimuli.

In these analyses, the optimal parameter value was calculated for each experiment. In what follows, all experiments will be analysed comprehensively using common parameters. The determination coefficient when parameters have fixed values will be calculated for each experiment along with their mean. The mean value to be calculated is the weighted average using Fisher's *Z* conversion. This procedure was repeated by changing the parameter values with 0.05 increment. Fig 1 shows the mean coefficient of determination for each pair of parameter values.

**Fig 1. Mean coefficient of determination for each pair of parameter values.** The mean value for the determination coefficient peaked when (α,m)=(0.35,-1.15) and the value was $\overline{R^2} = 0.93$.

Hattori and Oaksford [13] demonstrated that their *DFH* model (without parameters) showed the best fit compared to 33 models without parameters and seven models with parameters. We used eight types of experimental data to compare the performance of our proposed model with that of other models including $DFH$, $pARIs$, $\Delta P$ model, and conditional probability $P$. The results are shown in Table 3.

**Table 3. A replication trial of the meta-analysis of Hattori and Oaksford (13).**

| | AS95 | BCC03.1 | BCC03.3 | HO07.1 | HO07.2 | LS00 | W03.2 | W03.6 | $\overline{R^2}$ |
|---|---|---|---|---|---|---|---|---|---|
| $C$ | **0.92** | 0.90 | **0.97** | **0.97** | 0.96 | 0.72 | **0.91** | **0.93** | **0.93** |
| $P$ | 0.82 | 0.68 | 0.82 | 0.80 | 0.0 | 0.17 | 0.20 | 0.54 | 0.75 |
| $\Delta P$ | 0.78 | 0.84 | 0.70 | 0.50 | 0.0 | 0.77 | 0 | NA | 0.73 |
| $DFH$ | 0.91 | 0.95 | 0.91 | 0.93 | 0.96 | **0.80** | 0.69 | 0.80 | 0.91 |
| $pARIs$ | 0.89 | **0.96** | 0.94 | 0.93 | **0.99** | 0.80 | 0.78 | 0.88 | 0.90 |

Values in the table represent the determination coefficient between the estimated value of each model and the assessment value of the subject of each experiment. $C$ represents the extended confidence model when $\alpha$=0.35, m=-1.15. $P$ represents the conditional probability. The rightmost column represents the weighted average of the determination coefficient for each experiment. Bold-faced numbers represent the value of the model that marked the greatest determination coefficient for each experiment.

While the mean determination coefficient exceeded 0.9 for our proposed model, *DFH* and $pARIs$, it did not for the $\Delta P$ model or the conditional probability $P$. Particularly, the proposed model recorded the highest determination coefficient in five out of eight experiments, as well as in the mean of all experiments. Thus, it was shown that introducing symmetry bias into the conditional probability could significantly improve the determination coefficient with human assessment.

## Proposal of extended Bayesian inference

In this section, we propose the extended Bayesian inference where the conditional probability in Bayesian inference is replaced by the extended confidence model.

First, we describe Bayesian inference. This study deals with the problem of inferring a generative model (probability distribution) from observational data. To this end, in what follows, the hypothesis $h$ and data $d$ will be used on behalf of $p$ and $q$. Moreover, discrete models will be considered.

Bayesian inference first defines several hypotheses $h_i$ and provides a model for each hypothesis (probability distribution of data) in the form of conditional probability $\mathrm{P}(d \mid h_i)$. When data are fixed and regarded as a function of a hypothesis, this conditional probability is termed likelihood. The confidence $\mathrm{P}(h_i)$ for each hypothesis is given as a prior probability.

We can take $\mathrm{P}(d|h_i)$ and $\mathrm{P}(h_i)$ as initial values and calculate the posterior probability $\mathrm{P}(h_i|\boldsymbol{d})$ when observing data $\boldsymbol{d}$ using Bayes' theorem as follows.

$$P(h_i|\boldsymbol{d}) = P(h_i)\frac{P(\boldsymbol{d}|h_i)}{P(\boldsymbol{d})} \tag{3}$$

Hereinafter, data observed at a point in time are represented by the bold $\boldsymbol{d}$. Afterwards, we can replace the posterior probability with the prior probability using Bayesian updating.

$$P(h_i) \leftarrow P(h_i|\boldsymbol{d}) \tag{4}$$

By combining formulas (3) and (4), we get

$$P(h_i) \leftarrow P(h_i)\frac{P(\boldsymbol{d}|h_i)}{P(\boldsymbol{d})} \tag{5}$$

Whenever new data are observed, $P(h_i)$ in the formula (5), i.e., confidence in each hypothesis, is updated and the inference continues. The inference distribution during this procedure can be expressed as

$$P(d) = \sum_i P(h_i)P(d|h_i) \tag{6}$$

Note that in Bayesian inference, while the probability $P(h_i)$ of each hypothesis changes over time, the probability of the model of each hypothesis $P(d|h_i)$ does not. Fig 2 (a) shows an overview of the processing flow of Bayesian inference.

The extended Bayesian inference is an inference that has $\alpha$ and $m$ as parameters and accommodates normal Bayesian inference as its special case when $\alpha = 0$. Specifically, it is constructed by the following two update formulas.

$$C(h_i) \leftarrow \frac{C(h_i)C(\boldsymbol{d}|h_i)}{\left[(1-\alpha)(C(\boldsymbol{d}))^{-m} + \alpha(C(h_i))^{-m}\right]^{-1/m}} \tag{7}$$

$$C(\boldsymbol{d}|\boldsymbol{h}) \leftarrow \frac{C(\boldsymbol{h})C(\boldsymbol{d}|\boldsymbol{h})}{\left[(1-\alpha)(C(\boldsymbol{h}))^{-m} + \alpha(C(\boldsymbol{d}))^{-m}\right]^{-1/m}} \tag{8}$$

In formula (8), the bold-faced $\boldsymbol{h}$ represents the hypothesis that has the highest confidence. See the methods section below for a detailed derivation of the update formulas. Here, we can see that, supposing $\alpha = 0$ in formula (7), the right side shows the same form as that of Bayesian inference seen in formula (5).

$$C(h_i) \leftarrow \frac{C(h_i)C(\boldsymbol{d}|h_i)}{C(\boldsymbol{d})} \tag{9}$$

Now, in the case of Bayesian inference, the model $P(d|h_i)$ was invariable. We can ask if this is the same for extended Bayesian inference. Here, if we look closely to the right side of formula (8), supposing $\alpha = 0$, formula (8) becomes a tautology as shown below.

$$C(\boldsymbol{d}|\boldsymbol{h}) \leftarrow \frac{C(\boldsymbol{h})C(\boldsymbol{d}|\boldsymbol{h})}{C(\boldsymbol{h})} = C(\boldsymbol{d}|\boldsymbol{h}) \tag{10}$$

In other words, if $\alpha = 0$, then formula (8) substantially disappears. Conversely, if $\alpha > 0$, $C(\boldsymbol{d}|\boldsymbol{h})$ is subject to the denominator $C(\boldsymbol{d})$ in the right side, that is, the estimated value of the data. Following Gunji et al. [19], the process shown in formula (8) is termed Inverse Bayesian Inference.

In what follows we show the processing flow of Extended Bayesian inference. First, we take $P(d|h_i)$ and $P(h_i)$ as initial values and substitute them with $C$.

$$\begin{aligned} C(d|h_i) &= P(d|h_i) \\ C(h_i) &= P(h_i) \end{aligned} \tag{11}$$

Second, we calculate the degree of confidence for each hypothesis $C(h_i)$ and the model $C(\boldsymbol{d}|\boldsymbol{h})$ using the formula (7) and (8) whenever $\boldsymbol{d}$ is observed. Following the application of formulas (7) and (8), we can normalise $C(h_i)$ and $C(d|h_i)$ .

$$C(h_i) \leftarrow \frac{C(h_i)}{\sum_{k} C(h_k)}$$
$$C(d|h_i) \leftarrow \frac{C(d|h_i)}{\sum_{j} C(d_j|h_i)} \quad (12)$$

Finally, we can calculate the estimated distribution values as with Bayesian inference.

$$C(d) = \sum_{i} C(h_i)C(d|h_i) \quad (13)$$

Fig 2(b) shows an overview of the processing flow of extended Bayesian inference.

**Fig 2. A flow chart comparing Bayesian Inference to Extended Bayesian inference.** (a) An overview of the processing flow of Bayesian inference; and (b) An overview of the processing flow of extended Bayesian inference. In (b), for simplicity, the portion that belongs to the normalisation process is omitted. In (b), if we suppose α=0, the portion of inverse Bayesian inference disappears, corresponding to the Bayesian inference in (a).

**Performance evaluation of extended Bayesian inference using simulation**

To observe the behaviour of extended Bayesian inference, a simulation was performed. Specifically, a coin was tossed repeatedly, using a simulator, to observe the results and estimate the probability of getting heads using extended Bayesian inference. The probability of landing heads at the $t^{th}$ trial was designated $p^t$, and the probability of it landing tails was designated as $1 - p^t$ to handle cases where the probability changes over time. In each trial, a uniformly distributed random number was generated from interval [0.0, 1.0]. Numbers equal to or less than predefined $p^t$ were regarded as heads; numbers larger than $p^t$, were tails.

Whenever a coin toss result is observed, the correct probability of landing heads ($p^t$) is estimated by extended Bayesian inference. Additionally, for comparison with extended Bayesian inference, estimation

using only inverse Bayesian inference and estimation using Exponential Moving Average (EMA) were also carried out.

First, let heads be expressed as $HEAD$ and tails as $TAIL$. Second, we prepare $N$ hypotheses $(h_1, h_2, \cdots, h_N)$ and define the probability of heads and the probability of tails in each hypothesis $h_i$ as follows.

$$C(HEAD|h_i) = 0.5$$
$$C(TAIL|h_i) = 1.0 - C(HEAD|h_i) = 0.5 \quad (14)$$

That is, the models for all hypotheses are the same, and it means that this system has substantially no model of hypothesis at the initial stage.

Further, we must suppose that the prior probability for each hypothesis is equal.

$$C(h_i) = \frac{1}{N} \quad (15)$$

Whenever a coin toss result is observed, by performing extended Bayesian inference using formulas (7), (8), (12), and (13), $C(HEAD)$ is successively updated. In the simulations, $N = 3$. For the simplicity of subsequent analysis, in the following simulations, the parameter $m$ was fixed to $-1$ in the extended Bayesian inference.

When updating the degree of confidence $C(h_i)$ for each hypothesis using formula (7), we set the minimum value $\varepsilon$ to impose a restriction so that the degree of confidence will not be zero.

$$C(h_i) \leftarrow max\left(\varepsilon, \frac{C(h_i)C(\boldsymbol{d}|h_i)}{\left[(1-\alpha)(C(\boldsymbol{d}))^{-m} + \alpha(C(h_i))^{-m}\right]^{1/m}}\right) \quad (16)$$

Where $max\,(x,y)$ is a function whose output is a larger value of the two arguments $x,y$. In the simulation, $\varepsilon$ was set to 0.00001.

In case only inverse Bayesian inference is performed, the hypothesis is limited to only $\boldsymbol{h}$, the process of formula (7) is not performed, and $C(\boldsymbol{h})$ is always set to 1.0.

In this paper, we deal with a task in which the probability of heads can take two values, and they are replaced by the probability $\theta$. If a uniformly distributed random number generated from interval [0.0, 1.0] at the $t^{th}$ trial is denoted as $rnd^t$, the probability of heads is expressed by the following formula.

$$p^{t+1} = \begin{cases} p^t & if\ rnd^t > \theta \\ 1 - p^t & else \end{cases} \qquad (17)$$

In this simulation, $\theta$ was set to 0.0001. The initial value of the probability of heads $p^0$ was set to 0.85. That is, the probability of heads can take two values of 0.15 and 0.85.

EMA is calculated as a weighted average between the current estimated value and the observed data as follows.

$$s^{t+1} = \beta d^t + (1-\beta)s^t = s^t + \beta(d^t - s^t) \qquad (18)$$

Here, $s^t$ and $s^{t+1}$ represent estimated values of the probability of heads at $t^{th}$ and $t+1^{th}$ trial, respectively. $\alpha$ represents a learning rate, which takes a value of interval [0.0.1.0]. $d^t$ represents the result of coin toss at $t^{th}$ trial; in the case of $HEAD$, $d^t = 1$, and in the case of $TAIL$, $d^t = 0$. The weight of each data decreases exponentially as it goes to the past, and it is expressed by $\beta(1-\beta)^x d^{t-(x+1)}$. In the simulation, the value of $\beta$ was shifted from 0.0005 to 0.0063 with 0.0002 increment. That is, the estimations by EMA were performed using $\beta$ of the 30 patterns.

Fig 3 (a), (b), and (c) show the results of extended Bayesian inference, inverse Bayesian inference and estimations by EMA. However, for EMA, only three results of $\beta = 0.0005$, $\beta = 0.0021$ and $\beta = 0.0063$ are shown for easy viewing.

As can be seen from Fig 3(c), in the estimations by EMA, although rapid change can be followed as the learning rate $\alpha$ increases, the fluctuation in the stable period becomes larger. In other words, there is a trade-off between the ability to follow change and the accuracy in the stable period.

The result of only inverse Bayesian inference is very similar to the estimation result by EMA with $\beta = 0.0021$. On the other hand, although extended Bayesian inference takes time to follow changes as in the case of inverse Bayesian inference initially, the ability to follow gradually improves, and it becomes possible to respond rapidly to sudden changes.

**Fig 3. Time progress of the estimated values for the probability of head landing.** The figure includes the correct probability. (a) Estimated values by extended Bayesian inference. (b) Estimated values by only inverse Bayesian inference. (c) Estimated values by EMA.

Fig 4 shows the internal state of the extended Bayesian inference. Fig 4 (a) shows the time progress of the probability of head landing for each hypothesis. Initially, the probabilities for all hypotheses were 0.5 by definition, but learning by inverse Bayesian inference gradually formed hypothesis models. After the middle stage, the probabilities of head landing for three hypotheses $h_1$, $h_2$ and $h_3$ became approximately 0.15, 0.85, and 0.5, respectively. Here, 0.15 and 0.85 correspond to two correct values in this simulation, as shown in formula (17).

Figure 4(b) shows the time progress of the hypothesis with the greatest degree of confidence. As can be observed from the figure, in the second half of the simulation, extended Bayesian inference switches the hypotheses quickly when the probability of head landing is replaced. That is, extended Bayesian inference tries to respond to changes by learning using inverse Bayesian inference in the first half of the simulation, while in the second half of the simulation, abrupt changes are dealt with by switching the hypotheses formed by the learning.

**Fig 4. Internal state of extended Bayesian inference.** (a) Time progress of the probability of head landing for each hypothesis . (b) Time progress of the hypothesis with the greatest degree of confidence.

We show the relationship between the ability to follow the sudden change and the accuracy of the estimation in the stable period. Fig 5 shows the results of estimation by extended Bayesian inference, inverse Bayesian inference, and EMA in an enlarged manner between $40914^{th}$ trial at which the probability of heads suddenly changed from 0.15 to 0.85 and the $60913^{th}$ trial.

**Fig 5. Time progress of the estimated values for the probability of head landing.** The figure includes the correct probability.

This period was divided into the first half interval from $40914^{th}$ trial to $50913^{th}$ trial and the second half interval from $50914^{th}$ trial to $60913^{th}$ trial, and the differences between the correct values and the estimated values were calculated using root-mean-square error (RMSE) in each interval. RMSE is defined as follows.

$$RMSE = \sqrt{\frac{\sum_{t=k}^{T+k-1} (\hat{x}_t - x_t)^2}{T}} \qquad (19)$$

Here, $\hat{x}_t$ and $x_t$ represent the correct value and the estimated value in $t^{th}$ trial, respectively. $T$ represents the length of the interval.

We use the RMSE of the first half as a measure of the ability to follow rapid changes, and the RMSE of the second half as a measure of the accuracy of the estimation in the stable period.

Fig 6 shows the relationship between the followability and the estimation accuracy in the extended Bayesian inference, the inverse Bayesian inference, and EMA estimations.

As can be seen from the figure, there was a trade-off relationship of the accuracy being lost if the followability improved in EMA estimation. The regression curve for EMA data is also shown in this figure.

The data of the inverse Bayesian inference was located slightly lower left on this trade-off curve. On the other hand, the data of the extended Bayesian inference was almost the same as the data of the inverse Bayesian inference with regards to the accuracy, but the followability was greatly improved.

That is, it can be seen that the extended Bayesian inference broke the trade-off found in EMA estimation.

**Fig 6. Relationship between the followability and the estimation accuracy in the extended Bayesian inference, the inverse Bayesian inference, and EMA estimations.**

In the extended Bayesian inference, the inverse Bayesian inference could be applied only to the hypothesis $\boldsymbol{h}$ which has the greatest degree of confidence. Moreover, we set $\mathrm{m} = -1$. Because of this, we can rewrite formula (8) for inverse Bayesian inference as follows.

$$C(\boldsymbol{d}|\boldsymbol{h}) \leftarrow \frac{C(\boldsymbol{h})}{(1-\alpha)C(\boldsymbol{h}) + \alpha C(\boldsymbol{d})} \cdot C(\boldsymbol{d}|\boldsymbol{h}) \qquad (20)$$

Here, we can see that the denominator on the right side is the weighted average of $C(\boldsymbol{h})$ and $C(\boldsymbol{d})$, if $C(\boldsymbol{h}) > C(\boldsymbol{d})$, $C(\boldsymbol{h}) > (1-\alpha)C(\boldsymbol{h}) + \alpha C(\boldsymbol{d})$, so $C(\boldsymbol{d}|\boldsymbol{h})$ increases. At this point, the increment of $C(\boldsymbol{d}|\boldsymbol{h})$ is larger if the degree of confidence $C(\boldsymbol{h})$ is higher.

Conversely, if $C(\boldsymbol{h}) < C(\boldsymbol{d})$, $C(\boldsymbol{h}) < (1-\alpha)C(\boldsymbol{h}) + \alpha C(\boldsymbol{d})$, so $C(\boldsymbol{d}|\boldsymbol{h})$ decreases greatly if the degree of confidence $C(\boldsymbol{h})$ is lower.

Let us turn to the analysis of the stable period where $C(\boldsymbol{h}) = 1$. When $C(\boldsymbol{h}) = 1$, the total sum of confidence of all hypotheses is 1, and for any hypothesis $h_i$ other than $\boldsymbol{h}$, $C(h_i) = 0$. Hence, formula (13) can be rewritten as follows.

$$C(d) = \sum_i C(h_i)C(d|h_i) = C(d|\boldsymbol{h}) \qquad (21)$$

At this step, formula (20) for inverse Bayesian inference can also be rewritten as follows.

$$C(\boldsymbol{d}|\boldsymbol{h}) \leftarrow \frac{1}{(1-\alpha)\cdot 1 + \alpha C(\boldsymbol{d}|\boldsymbol{h})} \cdot C(\boldsymbol{d}|\boldsymbol{h}) \qquad (22)$$

The right side of this formula shows the weighted harmonic average of 1 and $C(\boldsymbol{d}|\boldsymbol{h})$. Since $0 \leq C(\boldsymbol{d}|\boldsymbol{h}) \leq 1$, the denominator is necessarily less than 1, and the likelihood $C(\boldsymbol{d}|\boldsymbol{h})$ increases whenever updated. In other words, when certain data are observed, the connection between the data and the hypothesis with the highest degree of confidence at that time is reinforced. Conversely, unobserved data, i.e., for $d_j$ other than $\boldsymbol{d}$, $C(d_j|\boldsymbol{h})$ can be standardised using formula (12), hence decreasing by the increment amount in $C(\boldsymbol{d}|\boldsymbol{h})$. Here, the rate of increase for $C(\boldsymbol{d}|\boldsymbol{h})$ depends on $\alpha$. When $\alpha = 1$, regardless of the presence value of $C(\boldsymbol{d}|\boldsymbol{h})$, the right side of formula (22) is 1. As $\alpha$ gets smaller, the increase rate lowers, and when $\alpha = 0$, it coincides with Bayesian inference and $C(\boldsymbol{d}|\boldsymbol{h})$ becomes invariable.

These considerations suggest that $\alpha$ becomes larger according to increase of updates to the model. In this sense, we can say that formula (20) for inverse Bayesian inference during the steady period represents a process of learning, and $\alpha$ corresponds to the rate of learning.

With respect to the portion that corresponds to Bayesian inference, suppose $\mathrm{m} = -1$ in the formula (7), then we can rewrite it as:

$$C(h_i) \leftarrow \frac{C(h_i)C(\boldsymbol{d}|h_i)}{\left[(1-\alpha)(C(\boldsymbol{d}))^{-m} + \alpha(C(h_i))^{-m}\right]^{-1/m}} = \frac{C(h_i)C(\boldsymbol{d}|h_i)}{(1-\alpha)C(\boldsymbol{d}) + \alpha C(h_i)} \quad (23)$$

Through careful observation of this formula we can note that $\alpha$ becomes larger, while $C(\boldsymbol{d})$, i.e., the effect of observation data, gets smaller. Where $\alpha = 1$, the denominator $C(\boldsymbol{d})$ disappears so $C(h_i)$ in the numerator and denominator is cancelled, and the formula can be expressed as follows.

$$C(h_i) \leftarrow C(\boldsymbol{d}|h_i) \quad (24)$$

This means that when $\alpha = 1$, the degree of confidence in each hypothesis $C(h_i)$ does not even consider the past observation history and is seen to be identical to the likelihood at that point in time. This coincides with the maximum likelihood estimation. In contrast, when $\alpha = 0$, the present formula coincides with the Bayesian inference expressed in formula (5).

A comparison of formula (5) with formula (24) reveals that their difference lies in the presence or absence of $C(h_i)$ on the right side, because $P(\boldsymbol{d})$ can be regarded as a constant. In this sense, the difference between them can be said to lie in the extent to which they accept history in order to determine degree of confidence.

As shown above, in extended Bayesian inference, symmetry bias plays two roles. First, the strength of symmetry bias indicates the rate of learning in portions of inverse Bayesian inference. In other words, the stronger the symmetry bias, the greater the degree of modification to the hypothesis model based on observational data. Second, the strength of the symmetry bias indicates how much the model takes into account past history in portions of Bayesian inference. In other words, as symmetry bias gets stronger, confidence in each hypothesis is updated based solely on more recent observational data.

## Discussion

In this study, we first proposed a different causal induction model. This model can replicate human judgments concerning causal induction with higher accuracy than previous models. Then we formulated an inference model that incorporates the said causal induction model into Bayesian inference. We noticed this inference model necessarily involves inverse Bayesian inference, which allows for flexibility to handle unsteady situations where the inference target changes from time to time. Finally, we demonstrated how this model can work well with unknown situations by forming new hypotheses through inverse Bayesian inference.

The causal induction model that we are proposing has two parameters that control the strength of symmetry bias. Conditional probability and causal induction models like *DFH* and *pARIs* can be shown to be special cases where particular values are assigned to parameters in our model. In other words, using the proposed model allows us to seamlessly express degree of confidence in those statements of the forms ‘if P then Q’, which stand for prediction, as well as those in the form of ‘P therefore Q’, which stand for causal

relation in a single model. The results of the meta-analysis of causal induction experiments revealed that the proper incorporation of symmetry bias in the proposed model allows it to replicate human judgment with high accuracy. However, as shown in Table 2, the values of the parameters are different for each experiment. It is known that the interpretation of conditionals largely change depending on the type and the contents of the conditionals, as well as subject's age [26]. Further studies are necessary to determine how parameters change according to type and contents of conditionals, as well as age.

The performance of the present model and the catalogue performance of *DFH* and *pARIs* were compared using eight sorts of experiment data. It turned out that the proposed model recorded the best determination coefficient in five of these experiments and the mean of all experiments. Thus, the present model is important in two ways. First, regarding human causal induction judgments, it has a capability that outperforms *DFH* and *pARIs*, which had hitherto demonstrated the best catalogue performance. Second, by using the extended confidence model, it is possible to determine the parameters $\alpha$ and $m$ that can best explain the participant's judgement from the data of simple causal induction experiments. In other words, we can measure the strength of symmetry bias the participant has. It would be possible to compare the strength of symmetry bias in patients with a mental illness against that of healthy individuals. For example, 'the Von Domarus principle' applies to the speech of schizophrenic patients [27] and refers to an inference of the form 'Men die. Weed die. Therefore, men are weed'. There is a widely observed tendency in schizophrenic patients to identify two things as the same when they share a common property – a mechanism said to underlie delusion [28]. Logically, it is wrong to conclude from 'A is C' and 'B is C' that 'A is B'. However, if the symmetry bias allows derivation of 'C is B' from 'B is C', then from 'A is C' and 'C is B', we can derive 'A is B'. In other words, one influence on the delusion of schizophrenic patients may be a strong susceptibility to symmetry bias. To test this hypothesis, it is possible to use the proposed model to estimate and compare the strength of symmetry bias in both patients with schizophrenia and healthy people. This is a research goal we can trace.

Parameters that denote the strength of symmetry bias indicate, in the case of extended Bayesian inference, the strength of inverse Bayesian inference, that is, the rate of learning. At the same time, they indicate how much the history is taken into account when updating the degree of confidence in each hypothesis. In this way, learning and inference become interlocked via parameters that denote the strength of symmetry bias in such way that it takes account only of more recent observational data in inference as the rate of learning becomes greater.

As a third form of inference following induction and deduction, various authors mention abduction. Abduction is an inference from knowledge or from known theories, for example, that 'if it rains, people open an umbrella' and to determine that from 'people open an umbrella' we can infer 'it is raining'. Abduction can be seen as a procedure of selecting hypothesis that best explain the observational data. In this regard, abduction is akin to maximum likelihood estimation and Bayesian inference. They differ, however, in that while the latter two inferences proceed by extracting the optimal hypothesis from the existing ones based on observational data, abduction focuses in the formation of a new hypothesis. An example is Newton, who introduced the law of universal gravitation to explain free fall of physical bodies.

Whereas the models for maximum likelihood estimation and Bayesian inference remain constant, the model for extended Bayesian inference is modified by virtue of inverse Bayesian inference enabling the model to match observation data. In other words, a new hypothesis that better explains the fact is formed in each case. In this sense, extended Bayesian inference that accompanies inverse Bayesian inference can be said to be akin to abduction.

In interpersonal communication, it is important to mutually estimate emotions of others. Further, for future studies on human-machine interaction this sort of information is essential. When estimating the emotion of others, since we cannot directly observe private internal states, there is no way to estimate their emotion other than using external clues (observational data) such as facial expression and tone of voice. In general, to perform proper estimation under these circumstances, the more observational data the better. However, emotions are not always constant, it is a variable that changes from time to time. Under these

467 circumstances, we must infer emotions while considering whether observational data to be used for
468 estimation derive from the same emotions. In the future it would be interesting to apply extended Bayesian
469 inference to tasks like these.

470 Suppose that we have knowledge (based on models of others) of the kinds of emotions the other has
471 and how they are expressed in him. Of course, one cannot attain perfect or complete knowledge because his
472 mental states have some degree of privacy. Suppose that when estimating based on incomplete knowledge
473 that the other is pleased, his expression suddenly changes. At this moment, one may think that his emotion
474 has changed or that this is another way of expressing pleasure. The former is an inference based on knowledge
475 and corresponds to Bayesian inference. The latter, on the other hand, is a modification of knowledge or an
476 addition of new knowledge and it corresponds to inverse Bayesian inference.

477 Incorporating the function of inverse Bayesian inference may help to develop robots that
478 autonomously learn and make various human-like inferences.

479

480 # Methods

481 **Data used in meta-analysis**

482 To test the descriptive performance of the extended confidence model, we performed the meta-
483 analysis using the same data as [13]. The analysis was conducted using eight types of experiment data, that
484 is, AS95, BCC03.1, BCC03.3, HO07.1, HO07.2, LS00, W03.2, and W03.6.

485 In AS95, forty graduate and undergraduate students were recruited. They were given co-occurrence
486 information about the presence or absence of drug treatment and the presence or absence of the side effects.
487 The subjects were asked to judge a number of problems, and each problem involved a sequence of instances
488 of these four information types. The frequencies of each information type varied from problem to problem.
489 At the end of a problem, the subjects were asked to enter a number from 0 to 100 that best reflected their
490 judgment of the drugs causing the side effects.

In BCC03.1, 109 undergraduate students were recruited and divided into two groups (preventive group and generative group). In preventive group, they evaluated how effectively each vaccine prevented the corresponding disease by giving a rating on a scale from 0 (the vaccine does not prevent the disease at all) to 100 (the vaccine prevents the disease every time). They also evaluated the influence of ray exposure on the mutation of viruses in generative group.

Thirty-one undergraduate students participated in BCC03.3. With regard to the side effects of drugs that reduce allergy, participants determined whether there were side effects of headache and, if so, assessed the causal strength between the drug and the headache.

In HO07.1, participants were 39 undergraduate students. They were asked to assess the strength of the causal relation between a particular type of fertiliser and plants blooming. They only observed a sequence of scenes in which fertiliser and plant blooming were either present or absent. After observing a series of situations, participants rated the subjective strength of the causal relation with a value between 0 (completely unrelated) and 100 (completely related).

In HO07.2, participants were 50 undergraduate students. In this experiment the cause was ‘drinking milk’ and the effect was ‘stomach-ache’. They judged the causal strength between drinking milk and stomach-ache according to given co-occurrence information.

In LS00, the participants of experiments 1, 2 and 3 were 27, 16, and 24 students, respectively. They assessed the extent to which a certain chemical causes a mutation in animals’ DNA using a number from between 0 and 100, where 0 indicates that the chemical does not cause mutations at all and 100 indicates that the chemical causes a mutation.

In W03.2, the participants were 40 undergraduate students. They were given information on the additives (manganese trioxide) contained in the foods a patient has eaten, and information on whether the patient has developed an allergic reaction. They were asked to judge the extent to which the statement ‘Manganese trioxide causes the allergic reaction in this patient’ was right for that patient and to write a

number from 0 (zero) to 100, where 0 (zero) means that the statement is definitely not right, and 100 means that the statement is definitely right.

In W03.6. the participants were 43 first-year undergraduate students. Most features of method, including initial written instructions; format of stimulus presentations; and procedure, were the same as in W03.2. The studies differed in design, however.

**Extended Bayesian inference**

First, we can replace $p$ and $q$ in formula (1) with $h_i$ and $d$.

$$C(d|h_i) = \left[(1-\alpha)P(d|h_i)^m + \alpha P(h_i|d)^m\right]^{1/m} \qquad (25)$$

Then we apply the Bayes' theorem to the right side of the formula (25), and we can obtain the conversion as follows.

$$\begin{aligned} C(d|h_i) &= \left[(1-\alpha)\left(\frac{P(h_i)P(d|h_i)}{P(h_i)}\right)^m + \alpha\left(\frac{P(h_i)P(d|h_i)}{P(d)}\right)^m\right]^{1/m} = \left[(1-\alpha)\left(\frac{1}{P(h_i)}\right)^m + \alpha\left(\frac{1}{P(d)}\right)^m\right]^{1/m} P(h_i)P(d|h_i) \\ &= \left[(1-\alpha)(P(h_i))^{-m} + \alpha(P(d))^{-m}\right]^{1/m} P(h_i)P(d|h_i) = \frac{P(h_i)P(d|h_i)}{\left[(1-\alpha)(P(h_i))^{-m} + \alpha(P(d))^{-m}\right]^{-1/m}} \end{aligned} \qquad (26)$$

By replacing $h_i$ and $d$ in formula (25), to perform the same conversion, we get

$$C(h_i|d) = \frac{P(d)P(h_i|d)}{\left[(1-\alpha)(P(d))^{-m} + \alpha(P(h_i))^{-m}\right]^{-1/m}} = \frac{P(h_i)P(d|h_i)}{\left[(1-\alpha)(P(d))^{-m} + \alpha(P(h_i))^{-m}\right]^{-1/m}} \qquad (27)$$

In the next step, we replace the conditional probability $P$ on the right side of formulas (26) and (27) with the extended confidence $C$ to make the formulas recursive, and then we replace the equation with the update formula.

$$C(d|h_i) \leftarrow \frac{C(h_i)C(d|h_i)}{\left[(1-\alpha)(C(h_i))^{-m} + \alpha(C(d))^{-m}\right]^{-1/m}} \qquad (28)$$

$$C(h_i|d) \leftarrow \frac{C(h_i)C(d|h_i)}{\left[(1-\alpha)(C(d))^{-m} + \alpha(C(h_i))^{-m}\right]^{-1/m}} \qquad (29)$$

As seen in formula (28), in inverse Bayesian inference, the amount of modification to the model of each hypothesis increases as $\alpha$ becomes larger. However, in this article, not all hypothetical models are uniformly modified, and the amount of modification changes according to confidence levels as follows.

$$\alpha_i = \alpha \cdot \pi(h_i) \tag{30}$$

However,

$$\pi(h_i) = \frac{exp\left(C(h_i)/\tau\right)}{\sum_{\boldsymbol{i}} exp\left(C(h_i)/\tau\right)} \tag{31}$$

Here, formula (31) is a procedure from the field of machine learning called Softmax [29], and $\tau$ $(>0)$ is a parameter termed temperature. $\pi$ remains the same value for all hypotheses if the temperature is high with the limit $\tau\to\infty$. On the other hand, if the temperature is low, $\pi$ becomes greater for hypotheses with a higher confidence level. $\pi$ takes the value 1 in the limit $\tau\to 0$ for hypotheses with the highest confidence level, and takes value 0 for all the other hypotheses.

In inverse Bayesian inference, the hypothetical model is modified when $\boldsymbol{d}$ is observed using $\alpha_i$ as follows.

$$C(\boldsymbol{d}|h_i) \leftarrow \frac{C(h_i)C(\boldsymbol{d}|h_i)}{\left[(1-\alpha_i)(C(h_i))^{-m} + \alpha_i(C(\boldsymbol{d}))^{-m}\right]^{-1/m}} \tag{32}$$

Here, there are reasons why the degree of modification for each hypothetical model changes according to the level of confidence. First, this process is a modification of the hypothetical model, which can be understood as a learning procedure rather than inference. Second, it is more likely that the currently observed data $\boldsymbol{d}$ derives from a hypothesis, if that hypothesis has a higher degree of confidence. Therefore, when modifying the model for each hypothesis based on observed data $\boldsymbol{d}$, the hypothesis with a higher degree of confidence requires a greater modification of its model. Of course, when $\tau\to\infty$, all hypothetical models can equally be modified. On the other hand, when $\tau\to 0$, only the hypothesis model with the highest

549 degree of confidence is modified. Moreover, supposing $\alpha = 0$, in formula (30), $\alpha_i = 0$ obtains for all
550 hypotheses.

551 In the simulation, $\tau \to 0$ was set. In other words, the inverse Bayesian inference was applied to only
552 the hypothesis $\boldsymbol{h}$ that has the highest confidence value.

553

554 ## Acknowledgements

555 This research is supported by the Center of Innovation Program from the Japan Science and Technology
556 Agency, JST and by JSPS KAKENHI Grant Numbers JP16K01408 and JP15H03002. We would like to
557 thank Editage [http://www.editage.com] for editing and reviewing this manuscript for English language.

558

559
560 ## References

561 1. Sidman M, Rauzin R, Lazar R, Cunningham S, Tailby W, Carrigan P. A search for symmetry in the
562 conditional discriminations of rhesus monkeys, baboons, and children. J Exp Anal Behav. 1982;37: 23-
563 44.
564
565 2. Sidman M, Tailby, W. Conditional discrimination vs. matching-to-sample: an expansion of the testing
566 paradigm. J Exp Anal Behav. 1982;37: 5-22.
567
568 3. Yamazaki Y. Logical and illogical behavior in nonhuman animals. Jap Psychol Res. 2004;46: 195-206.
569
570 4. Markman E, Wachtel G. Children's use of mutual exclusivity to constrain the meanings of words. Cogn
571 Psychol. 1988;20: 121-157.
572
573 5. Davidson D, Tell, D. Monolingual and bilingual children's use of mutual exclusivity in the naming of
574 whole objects. J Exp Child Psychol. 2005;92: 25-45.
575
576 6. Halberda, J. The development of a word-learning strategy. Cognition. 2003:87: B23-B34.
577
578 7. Lipkens R, Hayes S, Hayes L. Longitudinal study of the development of derived relations in an infant. J
579 Exp Child Psychol. 1993;56: 201-239.
580
581 8. O'Donnell J, Saunders K. Equivalence relations in individuals with language limitations and mental
582 retardation. J Exp Anal Behav. 2003;80: 131-157.
583

9. Takahashi T, Nakano M, Shinohara S. Cognitive symmetry: illogical but rational biases. Symmetry: Cult Sci. 2011;21: 275-294.

10. Kastak C, Schusterman R, Kastak D. Equivalence classification by California sea lions using class-specific reinforcers. J Exp Anal Behav. 2001;76: 131-158.

11. Tomonaga M, Matsuzawa T, Fujita K, Yamamoto J. Emergence of symmetry in a visual conditional discrimination by chimpanzees (Pan troglodytes). Psycholo Rep. 1991;68: 51-60.

12. Evans J, Handley S, Over D. Conditionals and conditional probability. J Exp Psychol: Learn mem cogn. 2003;29: 331-335.

13. Hattori M, Oaksford M. Adaptive non-interventional heuristics for covariation detection in causal induction: model comparison and rational analysis. Cogn Sci. 2007;31: 765-814.

14. Jenkins H, Ward W. Judgment of contingency between responses and outcomes. Psychol Monogr. 1965;79: 1-17.

15. Takahashi T, Oyo K, Tamatsukuri A. Correlation detection with and without the theory of conditionals: a model update of Hattori & Oaksford (2007). doi: 10.1101/247742.

16. Dehaene S. Consciousness and the brain: Deciphering how the brain codes our thoughts. New York: Viking; 2014.

17. Chater N, Oaksford M. The probabilistic mind: Prospects for Bayesian cognitive science. Oxford: Oxford University Press; 2008.

18. Arecchi F. Phenomenology of consciousness from apprehension to judgment: nonlinear dynamics. Psychol Life Sci. 2011;15: 359-375.

19. Gunji Y, Shinohara S, Haruna T, Basios V. Inverse Bayesian inference as a key of consciousness featuring a macroscopic quantum logical structure. Biosystems. 2016;152: 44-65.

20. Gunji Y, Murakami H, Tomaru T, Basios V. Inverse Bayesian inference in swarming behaviour of soldier crabs. Philos Trans A Math Phys Eng Sci. 2018;376: 20170370.

21. Horry Y, Yoshinari A, Nakamoto Y, Gunji Y. Modeling of decision-making process for moving straight using inverse Bayesian inference. Biosystems. 2018;163: 70-8.

22. Anderson J, Sheu C. Causal inferences as perceptual judgments. Mem Cogn. 1995;23: 510-524.

23. Buehner M, Cheng P, Clifford D. From covariation to causation: a test of the assumption of causal power. J Exp Psychol Learn Mem Cogn. 2003;29: 1119-1140.

24. Lober K, Shanks D. Is causal induction based on causal power? Critique of Cheng (1997). Psychol Rev. 2000;107: 195-212.

25. White P. Making causal judgments from the proportion of confirming instances: the pCI rule. J Exp Psychol Learn Mem Cogn. 2003;29: 710-727.

26. Gauffroy C, Barrouillet P. Heuristic and analytic processes in mental models for conditionals: an integrative developmental theory. Dev Rev. 2009;29: 249-282.

27. Von Domarus E. The specific laws of logic in schizophrenia. In: Kasanin JS, editor. Language and thought in schizophrenia. Berkeley: University of California Press; 1944. pp. 104-113.

28. Arieti S. Interpretation of schizophrenia. New York: Basic Books; 1957.

29. Sutton R, Barto A. Reinforcement learning: An introduction. Cambridge, MA: MIT press; 1998.

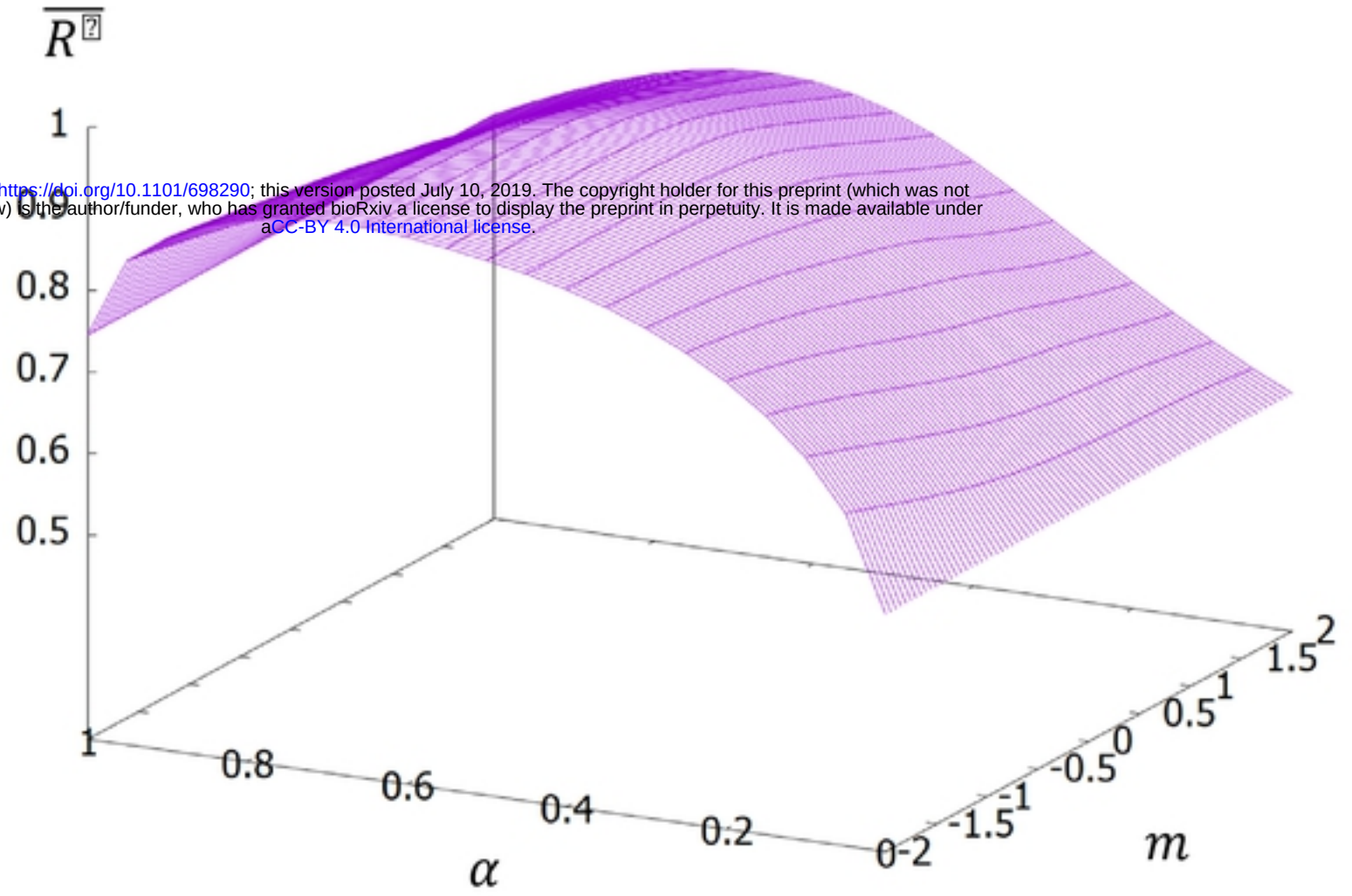
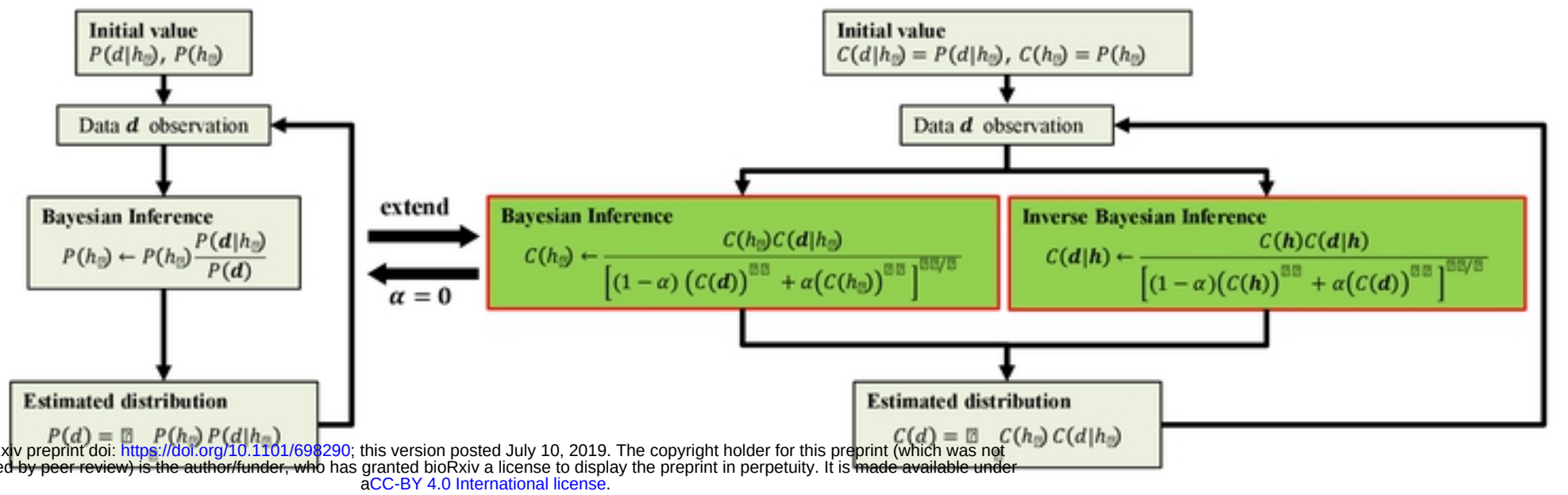


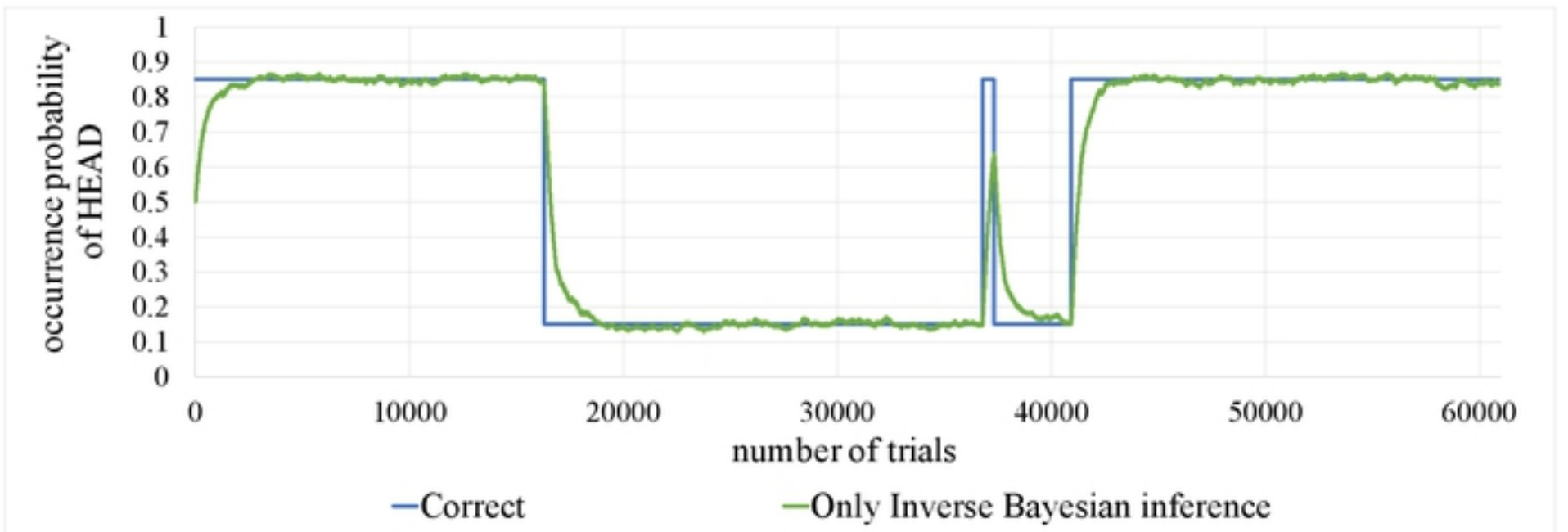

**Fig. 1.** Mean coefficient of determination for each pair of parameter values. The mean value for the determination coefficient peaked when $(\alpha,m) = (0.35, -1.15)$ and the value was $\overline{R^2} = 0.93$.



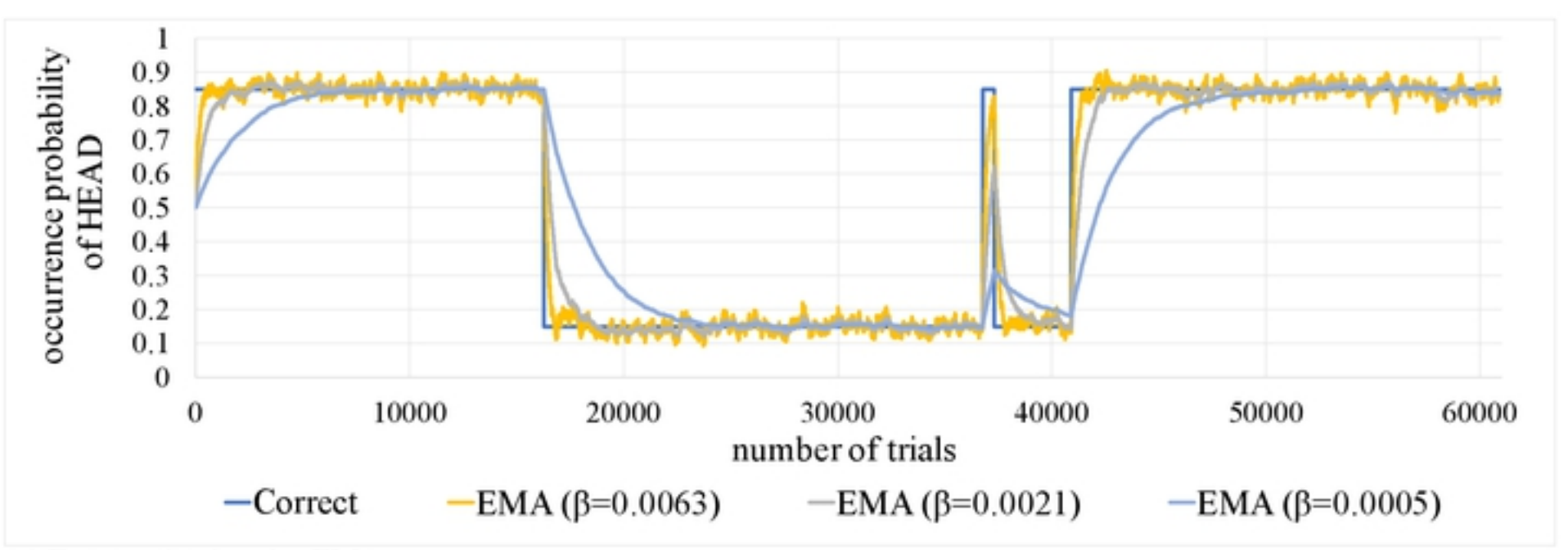

(a) Bayesian Inference (b) Extended Bayesian Inference

**Fig. 2.** A flow chart comparing Bayesian Inference to Extended Bayesian inference. (a) An overview of the processing flow of Bayesian inference; and (b) An overview of the processing flow of extended Bayesian inference. In (b), for simplicity, the portion that belongs to the normalisation process is omitted. In (b), if we suppose $\alpha = 0$, the portion of inverse Bayesian inference disappears, corresponding to the Bayesian inference in (a).

(a) Estimated values by extended Bayesian inference $(\alpha,m,\tau) = (0.01, -1.0, 0.0)$

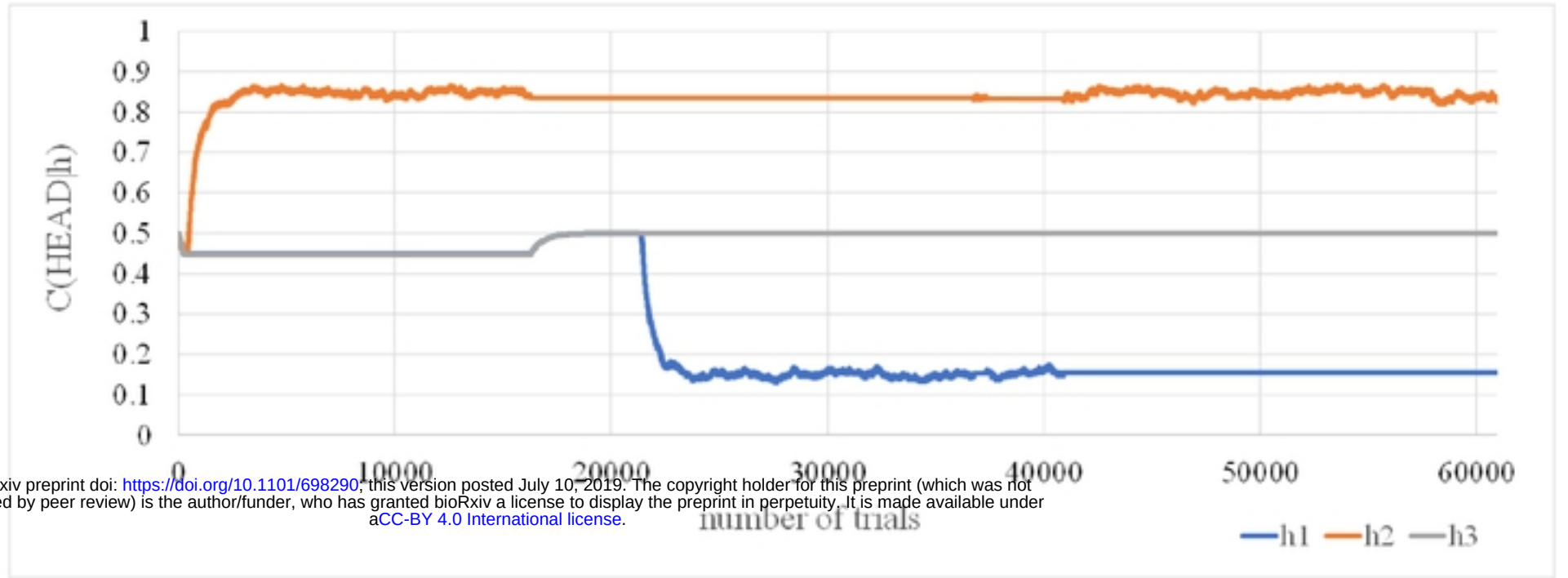



(b) Estimated values by only inverse Bayesian inference $(\alpha,m,\tau) = (0.01, -1.0, 0.0)$

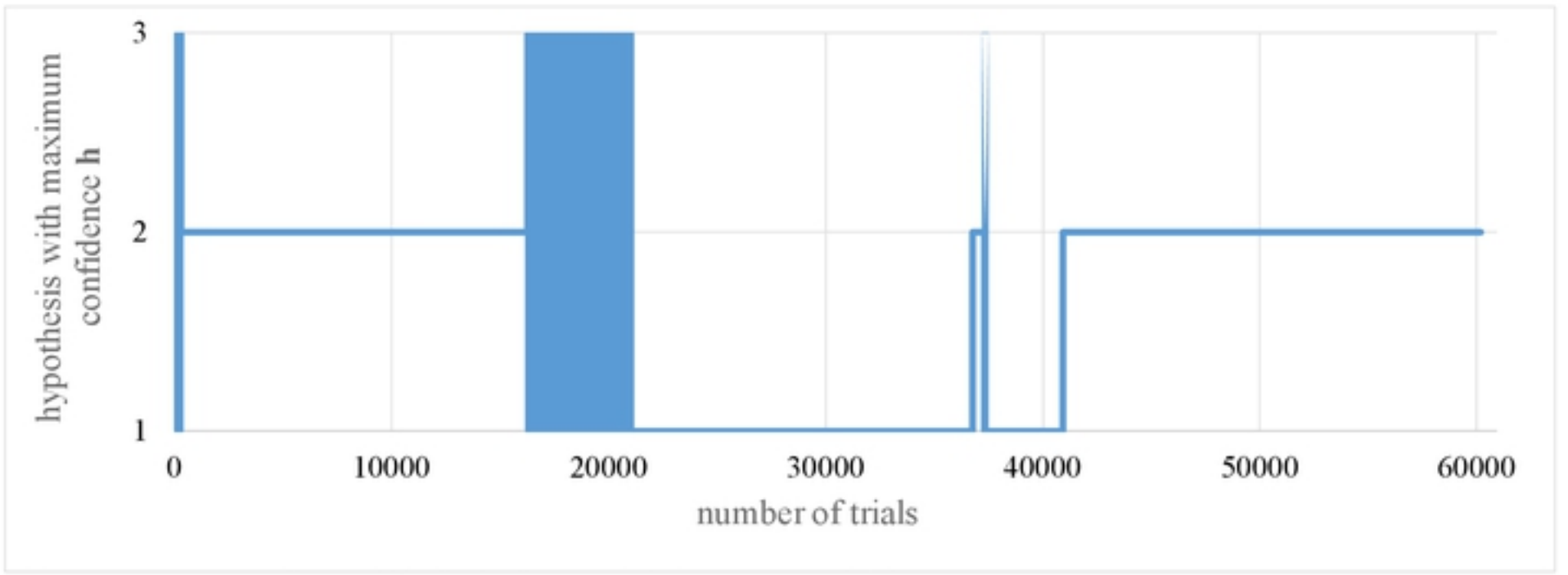



(c) Estimated values by EMA

**Fig. 3** Time progress of the estimated values for the probability of head landing. The figure includes the correct probability.




(a) Time progress of the probability of head landing for each hypothesis




(b) Time progress of the hypothesis with the greatest degree of confidence

**Fig. 4** Internal state of extended Bayesian inference

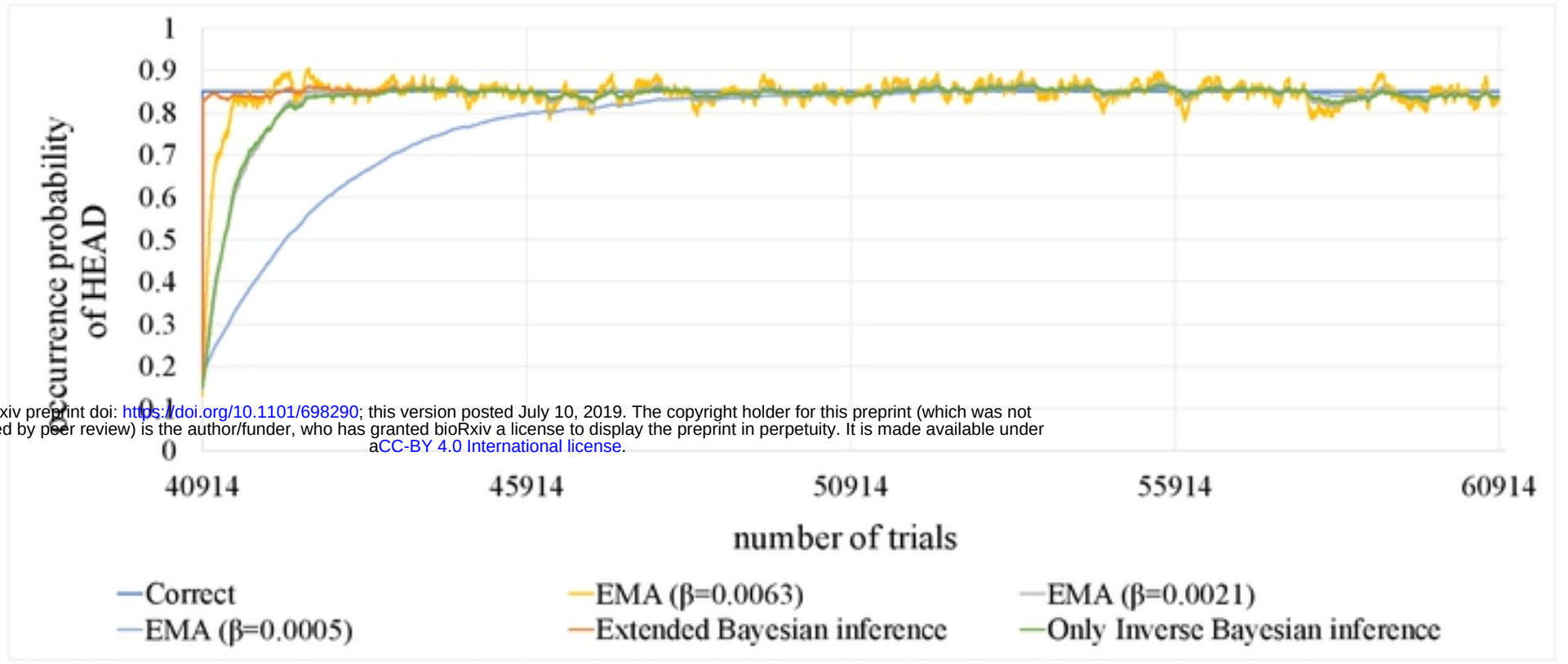



**Fig. 5** Time progress of the estimated values for the probability of head landing. The figure includes the correct probability.

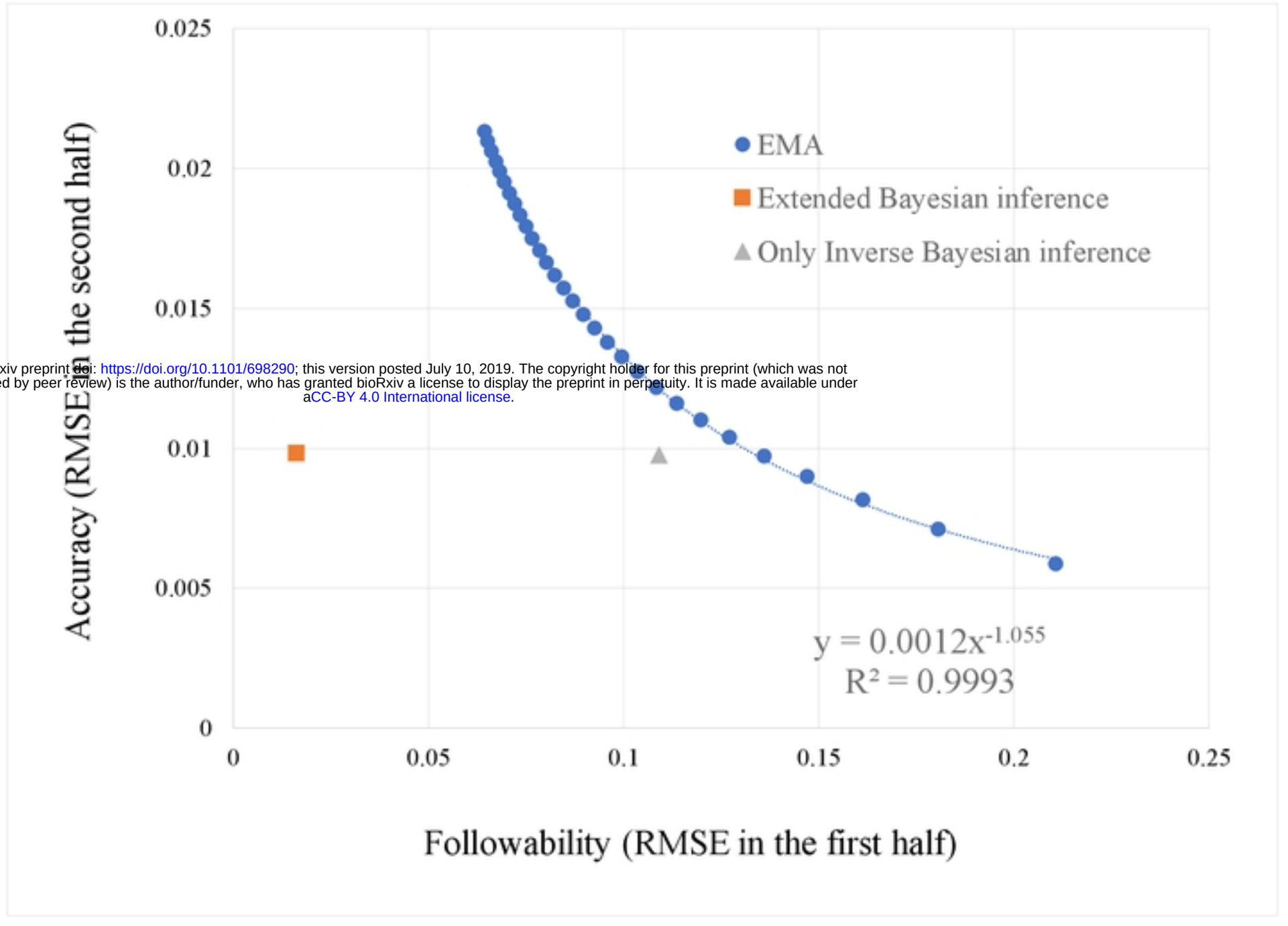



**Fig. 6** Relationship between the followability and the estimation accuracy in extended Bayesian inference, inverse Bayesian inference, and EMA estimations.