Decision confidence reflects maximum entropy reinforcement learning

Amelia M. Johnson
University of Washington
Seattle, WA 98195
Allen Institute
Seattle, WA 98109

Michael A. Buice
Allen Institute
Seattle, WA 98109
University of Washington
Seattle, WA 98195

Koosha Khalvati [*]
Allen Institute
Seattle, WA 98109

Abstract

Current computational models have not been able to account for the effect of reward in confidence reports among humans. Here we propose a mathematical framework of confidence that is able to generalize across various decision making tasks involving varying prior and reward distributions. This framework proposes a formal definition of "decision confidence" through the concept of soft optimality. We further show that the objective function in this framework is jointly maximising the reward and information entropy of the policy. We confirm the validity of our framework by testing it on data gathered under various task conditions.

1 Introduction

Self-assessment of one's choices, i.e. decision confidence, plays a key role in long-term decision making and learning [1]. This assessment helps the decision maker improve their model of the outside world and consequently gain higher utility in the future [2, 3]. Therefore, it is important to come to an understanding of how these confidence approximations are internally calculated and especially how given information about a decision making scenario is incorporated into these approximations.

Despite our general understanding of the word "confidence" in day-to-day life, posing a formal definition of confidence is very challenging. Specifically, confidence is only mathematically defined for scenarios where different choices have exactly the same amount of (potential) reward. In these situations the goal is to find and pick the "correct" choice as opposed to all other equally "incorrect" ones. Consequently, confidence is defined as the probability of choosing the correct option, which is basically the posterior probability of the most probable choice given the received sensory/cognitive information and the decision maker's prior knowledge of the environment [4, 1, 5, 6]. This definition, sometimes referred to as the "Bayesian confidence hypothesis", does not consider the utility of different choices, which is not required given the equality of these values.

In many real life situations, however, different actions lead to different utilities. This difference often plays a key role in the decision making process. A rustling sound is more likely to be because of the wind rather than a predator approaching, but everyone becomes more vigilant when hearing that sound as one of the possibilities is potentially life threatening despite being unlikely. Moreover, everyone would be certain about the rationality of this decision. The current definition of confidence is incapable of modeling this certainty/confidence.

Here we present a mathematical framework to formally define and assess confidence in a general decision making scenario involving uncertainty about the outside world, prior knowledge of the decision maker about the world, and different utility functions for available choices. In order to

[*] Emails: ameliamj@uw.edu, michaelbu@alleninstitute.org, koosha.khalvati@alleninstitute.org

Information-Theoretic Principles in Cognitive Systems Workshop at the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)

incorporate the utility function into the confidence calculation, we model confidence as "probability of making the best decision". Formally, confidence is defined as probability of being optimal over a sequence of states and actions given the policy.

We further show that our approach to modelling subjects' confidence in their decision equals to a planning as inference framework [7, 8]. This framework maps to a reinforcement learning agent whose objective function is to jointly maximize the reward and the information entropy of the policy (also called maximum entropy reinforcement learning) [9, 10]. We validated our normative approach by testing it on an open data set with various task conditions involving asymmetries in the task prior and choice rewards.

2 Background

A **Partially Observable Markov Decision Process (POMDP)** is a mathematical framework for planning and decision making under uncertainty in the field of Artificial Intelligence [11]. It is formally defined as a tuple (S, A, Z, P, O, R) where S is the set of states, A is the set of actions, Z is the set of observations, $P = p(s'|a, s)$ is the transition function between states, $O = p(z|s, a)$ is the observation function, and $R = r(s, a)$ is the reward function expressing the reward/utility of the agent given the state and action. A POMDP agent does not know the current state of the environment. Therefore, starting from a prior, called the initial belief state b_0, it updates the posterior probability distribution over the states with each observation and action:

$$b_t(s) \propto p(z_t|s, a_{t-1}) \sum_{s' \in S} p(s|s', a_{t-1}) b_{t-1}(s') \tag{1}$$

The goal of the POMDP agent is to gain maximum accumulated reward within the horizon H. The recipe for action selection, called *policy* π, can be represented as a mapping from belief states to actions with the optimal policy π^* being the mapping that maximize the expected total reward:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{(b_t, a_t)} \left[\sum_{t=0}^{H} \sum_{s \in S} b_t(s) r(s, a_t) \right] = \arg\max_{\pi} \mathbb{E}_{(b_t, a_t)} \left[\sum_{t=0}^{H} r(b_t, a_t) \right]. \tag{2}$$

Models of decision making in Cognitive Neuroscience: POMDPs and similar Bayesian approaches have been shown to be extremely successful in modelling behaviour of subjects across various decision making tasks and species [5, 12, 13]. One of the main applications of POMDPs is modelling "perceptual decision making" where the subject should select the "correct" choice based on some sensory observations to get reward [14]. As the term "correct" suggests, the reward function is symmetrical among different choices. In other words, all choices have the same "value". In these studies, the subject's confidence closely matches the probability of choosing the correct option, i.e. the posterior probability of the most probable choice [1, 6].

Currently, there is no normative model of confidence for situations where the reward value of options differs, which in fact, happens more frequently in real life compared to symmetric reward situations. There are some confidence experiments with asymmetries in the reward function. Models and methods of these studies, however, are all descriptive/statistical, e.g., positive correlation between confidence report and reward value [15, 16, 17]. Here we present a unifying framework about the "decision confidence" which models the confidence of the subject in their decision. In other words, we present the mathematical model for the following question the subject might ask themselves: "What is the probability that I made the best decision?"

3 Model

To formally model "the best decision", we include the idea of optimality to the POMDP framework, where optimality is a binary variable that is reflective of receiving the maximum reward. The probability of optimality is used to describe a subject's confidence (or their internal sense of whether the decision they made was optimal). Importantly, the decision making process might involve multiple actions. Therefore, what we refer to as a "decision" is in fact a sequence of actions, given an observation after each action (which updates the belief), shown by $\tau = b_0, a_0, b_1, a_1, \ldots, a_H$ and called a trajectory. Consequently, the probability of being optimal is $p(O = 1|\tau)$. This probability should be maximal for a trajectory that is generated by the optimal policy π^*.

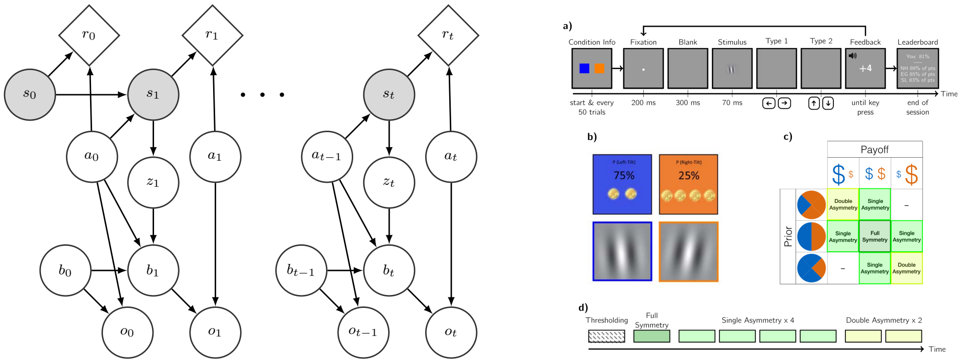

Figure 1: Left: Graphical model of the framework that measures the probability of making the optimal decision as confidence. Right: Experimental setup of a perceptual decision making task with varying prior and reward distribution (picture from [17]).

Given the fact that the system is Markovian and each policy is a mapping from belief states to actions, the probability of optimality can be expressed for each action given the belief state, i.e. $p(o_t = 1|a_t, b_t)$ (Fig. 1, left plot). With this representation the probability of optimality for the whole trajectory can be expressed as $p(o_{1:H} = 1|\tau)$. Consequently, the probability of a trajectory being optimal is:

$$p(o_{0:H} = 1|\tau = b_0, a_0, \ldots, a_H) = \prod_{t=0}^{H} p(o_t = 1|b_t, a_t) \tag{3}$$

The above distribution should reflect the expected accumulated reward, i.e. higher total reward should mean a higher probability of optimality. One straightforward way to satisfy this, is to set probability of optimality for each single action as following:

$$p(o_t = 1|b_t, a_t) \propto e^{r(b_t, a_t)} \tag{4}$$

Since many functions other than exponentiation might reflect the accumulation of reward (at least in specific cases), it is important to examine the driven policy from maximizing $p(\tau|o_{0:H})$ when we define the probability of being optimal with Eq. 4:

$$p(\tau|o_{0:H}) \propto p(\tau, o_{0:H}) = p(b_0) \prod_{t=0}^{H} p(b_{t+1}|b_t, a_t) p(o_t|b_t, a_t) = p(b_0) \prod_{t=0}^{H} p(b_{t+1}|b_t, a_t) e^{r(b_t, a_t)}. \tag{5}$$

One way to derive such a policy is to approximate $p(\tau|o_{0:H})$. If the approximation of the optimal policy $p(a_t|o_t = 1, b_t)$ is expressed with $\pi(a_t|b_t)$, the approximation of optimal trajectory will be:

$$\hat{p}(\tau) = p(b_0) \prod_{t=0}^{H} p(b_{t+1}|b_t, a_t) \pi(a_t|b_t). \tag{6}$$

and the desired policy $\pi(a_t|b_t)$ is obtained by minimizing the KL-divergence between the approximate (Eq. 6) and the true distribution (Eq. 5):

$$\begin{aligned} D_{KL}(\hat{p}(\tau)||p(\tau|o_{0:H})) &= -\mathbb{E}_{\tau \sim \hat{p}(\tau)} \left[\sum_{t=0}^{H} p(b_{t+1}|b_t, a_t) + r(b_t, a_t) - p(b_{t+1}|b_t, a_t) - \log \pi(a_t|b_t) \right] \\ &= -\sum_{t=0}^{H} \mathbb{E}_{(b_t, a_t) \sim \hat{p}(b_t, a_t)} \left[r(b_t, a_t) - \mathcal{H}(\pi(a_t|b_t)) \right] \end{aligned} \tag{7}$$

This means that an agent with this definition of optimality jointly maximizes (minimizes the negative of) the total reward and the information entropy of the policy, which is also soft Q-learning on the belief state [9, 10]. This derivation is known as "planning as inference" in the literature [7, 8]. The policy derived through this process is "soft optimal".

Table 1: AIC values by subject for different confidence prediction models

Model	1	2	3	4	5	6	7	8	9	10
Observation	1359	**1670**	1793	1293	**1439**	**1368**	1710	1737	1654	2413
Posterior	1418	1677	1751	1293	**1439**	1394	**1676**	2032	1791	1369
Expected Value	1414	1675	1687	**1289**	**1439**	1385	1680	1943	1730	2409
Decision	**1199**	1806	**1676**	2111	2347	1779	1924	**1666**	**1652**	**1357**

4 Results

We tested our model on a perceptual decision making experiment measuring confidence in different conditions including asymmetric priors and reward [17]. In this experiment, 10 subjects were shown a Gabor filter tilted left or right and asked to maximize their score by reporting the direction of the Gabor filter (left or right) and their confidence (low or high). Each subject completed 7 session of this task where the prior probability distribution and the reward (score) distribution for correction choices across the two direction was varied (prior probability was either 3:1, 1:1, or 1:3; rewarded score were either 2:4, 3:3, or 4:2). The subjects were told the exact prior and reward distribution before each session and every 50 trials during the session. Importantly, while subjects wanted to maximize their score, they were explicitly asked to report their perceptual confidence. This means that the reward was not supposed to be included in confidence report. Despite this explicit direction, confidence of many subjects were affected in sessions with asymmetric reward function as reported in the original paper of this study[17].

To approximate the generative function for perception of each subject, we fit a mixture of 2 Gaussian distributions, $\mathcal{N}(1, \sigma_z)$ and $\mathcal{N}(-1, \sigma_z)$ (1 free parameter), to their performance in the fully symmetric session. Then we fit another mixture, $\mathcal{N}(1, \sigma_{sz})$ and $\mathcal{N}(-1, \sigma_{sz})$ where σ_{sz}^2 represented sensory observation variance according to the subject's internal model, to the subject's performance in the session with asymmetric priors. Moreover, the confidence criterion, the threshold at which confidence is binarized into "low" or "high", was fit to the subjects' confidence reports in both fully symmetric and asymmetric prior sessions by using grid search and a maximum likelihood estimation with a Bernoulli likelihood function.

We compared the fit of our model, labelled "Decision" in Table 1, to subjects' behavioural data to the fit of 3 other baseline models in the Bayesian framework. These baseline models were the observation model, which predicts confidence as proportional to the likelihood of a subject's observation ($p(z|s)$); the posterior model, which predicts confidence as proportional to the posterior probability ($p(s|z)$); and the expected value model, which predicts confidence as proportional to the expected value ($E_{p(s|z)}[r(s)]$). Since this experiment includes only one action the confidence measured by our framework was basically the ratio of exponentiation of the expected reward. The parameters related to the performance, i.e. σ_z and σ_{sz}, were the same in all of the 4 models. On the other hand, the threshold that binarized the confidence were fitted separately for each confidence model.

Table 1 shows the AIC values of each model tested on the session with asymmetric reward. These values show that 5 out of 10 subjects reported their decision confidence. This is especially interesting since the subjects were explicitly asked to report the posterior belief (perception confidence) and ignore the reward in their confidence report. This phenomena suggests that in the self-assessment mechanism of the brain, prior and reward are not easily separable. Such inseparability itself is a signature of the planning as inference framework where all aspects of the environment are put together in the probability of being optimal.

5 Conclusion

Here we presented a generalizable framework that measures "decision confidence". This confidence is basically the probability of a trajectory being optimal, considering the probability distribution of the current state, and reward of each action. Our framework suggests that decision confidence reflects "soft optimality" which includes maximizing both reward and information entropy of the policy, also called maximum entropy reinforcement learning. We validated our framework by testing it on an experiment with varying prior and reward distributions.

References

[1] Alexandre Pouget, Jan Drugowitsch, and Adam Kepecs. Confidence and certainty: distinct probabilistic quantities for different goals. *Nature Neuroscience*, 19(3):366–374, March 2016. Number: 3 Publisher: Nature Publishing Group.

[2] Roozbeh Kiani and Michael N. Shadlen. Representation of Confidence Associated with a Decision by Neurons in the Parietal Cortex. *Science*, 324(5928):759–764, May 2009. Publisher: American Association for the Advancement of Science.

[3] Jan Drugowitsch, André G Mendonça, Zachary F Mainen, and Alexandre Pouget. Learning optimal decisions with confidence. *Proceedings of the National Academy of Sciences*, 116(49):24872–24880, 2019. Publisher: National Acad Sciences.

[4] Rubén Moreno-Bote. Decision Confidence and Uncertainty in Diffusion Models with Partially Correlated Neuronal Integrators. *Neural Computation*, 22(7):1786–1811, July 2010.

[5] Koosha Khalvati and Rajesh PN Rao. A Bayesian Framework for Modeling Confidence in Perceptual Decision Making. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[6] Koosha Khalvati, Roozbeh Kiani, and Rajesh P. N. Rao. Bayesian inference with incomplete knowledge explains perceptual confidence and its deviations from accuracy. *Nature Communications*, 12(1):5704, September 2021. Number: 1 Publisher: Nature Publishing Group.

[7] Marc Toussaint and Amos Storkey. Probabilistic inference for solving discrete and continuous state Markov Decision Processes. In *Proceedings of the 23rd international conference on Machine learning - ICML '06*, pages 945–952, Pittsburgh, Pennsylvania, 2006. ACM Press.

[8] Matthew Botvinick and Marc Toussaint. Planning as inference. *Trends in Cognitive Sciences*, 16(10):485–488, October 2012.

[9] Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd national conference on Artificial intelligence - Volume 3*, AAAI'08, pages 1433–1438, Chicago, Illinois, July 2008. AAAI Press.

[10] Sergey Levine. Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review, May 2018. arXiv:1805.00909 [cs, stat].

[11] Anthony R. Cassandra, Leslie Pack Kaelbling, and Michael L. Littman. Acting optimally in partially observable stochastic domains. In *Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence*, AAAI'94, pages 1023–1028, Seattle, Washington, August 1994. AAAI Press.

[12] Rajesh P. N. Rao. Decision Making Under Uncertainty: A Neural Model Based on Partially Observable Markov Decision Processes. *Frontiers in Computational Neuroscience*, 4:146, November 2010.

[13] Koosha Khalvati, Saghar Mirbagheri, Seongmin A. Park, Jean-Claude Dreher, and Rajesh PN Rao. A Bayesian Theory of Conformity in Collective Decision Making. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[14] Joshua I. Gold and Michael N. Shadlen. The Neural Basis of Decision Making. *Annual Review of Neuroscience*, 30(1):535–574, 2007. _eprint: https://doi.org/10.1146/annurev.neuro.29.051605.113038.

[15] Benedetto De Martino, Stephen M. Fleming, Neil Garrett, and Raymond J. Dolan. Confidence in value-based choice. *Nature Neuroscience*, 16(1):105–110, January 2013. Number: 1 Publisher: Nature Publishing Group.

[16] Annika Boldt, Charles Blundell, and Benedetto De Martino. Confidence modulates exploration and exploitation in value-based learning. *Neuroscience of Consciousness*, 2019(1):niz004, January 2019.

[17] Shannon M. Locke, Elon Gaffin-Cahn, Nadia Hosseinizaveh, Pascal Mamassian, and Michael S. Landy. Priors and payoffs in confidence judgments. *Attention, Perception, & Psychophysics*, 82(6):3158–3175, August 2020.