OPEN ACCESS Freely available online

PLOS ONE

Reward Maximization Justifies the Transition from Sensory Selection at Childhood to Sensory Integration at Adulthood

Pedram Daee[1]*, Maryam S. Mirian[1], Majid Nili Ahmadabadi[1,2]

1 Cognitive Robotics Laboratory, Control and Intelligent Processing Center of Excellence, School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran, **2** School of Cognitive Sciences, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran

Abstract

In a multisensory task, human adults integrate information from different sensory modalities -behaviorally in an optimal Bayesian fashion- while children mostly rely on a single sensor modality for decision making. The reason behind this change of behavior over age and the process behind learning the required statistics for optimal integration are still unclear and have not been justified by the conventional Bayesian modeling. We propose an interactive multisensory learning framework without making any prior assumptions about the sensory models. In this framework, learning in every modality and in their joint space is done in parallel using a single-step reinforcement learning method. A simple statistical test on confidence intervals on the mean of reward distributions is used to select the most informative source of information among the individual modalities and the joint space. Analyses of the method and the simulation results on a multimodal localization task show that the learning system autonomously starts with sensory selection and gradually switches to sensory integration. This is because, relying more on modalities -i.e. selection- at early learning steps (childhood) is more rewarding than favoring decisions learned in the joint space since, smaller state-space in modalities results in faster learning in every individual modality. In contrast, after gaining sufficient experiences (adulthood), the quality of learning in the joint space matures while learning in modalities suffers from insufficient accuracy due to perceptual aliasing. It results in tighter confidence interval for the joint space and consequently causes a smooth shift from selection to integration. It suggests that sensory selection and integration are emergent behavior and both are outputs of a single reward maximization process; i.e. the transition is not a preprogrammed phenomenon.

Citation: Daee P, Mirian MS, Ahmadabadi MN (2014) Reward Maximization Justifies the Transition from Sensory Selection at Childhood to Sensory Integration at Adulthood. PLoS ONE 9(7): e103143. doi:10.1371/journal.pone.0103143

Editor: Robert J. van Beers, VU University Amsterdam, Netherlands

Received March 19, 2014; **Accepted** June 27, 2014; **Published** July 24, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are within the paper.

Funding: The authors have no funding or support to report.

Competing Interests: The authors have declared that no competing interests exist.

* Email: pedram.daee@gmail.com

Introduction

To make an appropriate decision, our brain has to perceive the current state of the environment. However, even our best senses are noisy and can only provide an uncertain estimate of the underlying state. The biological solution for achieving the best perception is integration of uncertain individual estimates.

Human adults integrate sensory information, both across and within different modalities, with seemingly the purpose of reducing the uncertainty of their perception. The overwhelming majority of behavioral studies have shown that this uncertainty reduction happens in a statistically optimal fashion [1], [2]. One way to model this optimal integration is employing the Bayesian framework. In this framework and under some assumptions, the integration procedure is modeled by a weighted average of the individual sensors' estimates. Each sensor's weight is proportional to its relative reliability; i.e. inverse of its uncertainty. It can be shown that the reliability of the integrated estimate is higher than that of any individual's estimate.

Nevertheless, many behavioral studies indicate that this optimal behavior, and in some cases even its neural foundations, are not present at birth. Furthermore, it is only in the later stages of development that multisensory functions appear and take the main role in multisensory decision makings; see [3] for a comprehensive review. An increasing number of studies in different sensory modalities on adults and children have shown that, unlike adults, children make their judgments based only on one of the available sources of information. Some instances of this sensory selection behavior have been observed in visual and haptic modalities for size and orientation discrimination [4], visual landmarks and self-motion information for navigation [5], and visual stereoscopic and texture information for estimating surface slant [6].

The interesting open questions here are "Why does optimal integration occur so late?" [7], why there is a tendency in sensory selection in children, and finally, how and based on what measures does the transition from sensory selection at childhood to sensory integration at adulthood happen. While there are a considerable number of hypotheses regarding the reasons behind these phenomena (see [6], [3], [7]), to our knowledge, no existing study has addressed these three questions with a unified computational model. The primary aim of this research is to investigate the computational advantages of the transition from sensory selection

at early ages toward multisensory integration at adulthood. The second goal is to check if the above three questions can be addressed by a single computational model.

We hypothesize that this selection and integration are emergent behavior of a single reward maximization system. To verify our hypothesis, we propose a mathematically sound and general reward dependent learning framework (see Method) and test it in a multisensory localization task (see Experiments and Results). The learning method is value-based [8] [9] and progress of learning in the framework corresponds to development of the agent over age. This choice is natural as there are supporting studies indicating that the multisensory integration is not innate and there should be a learning mechanism behind its development (see [3], [10]). Furthermore, this framework does not require most of the strict mathematical assumptions that are building blocks of the conventional Bayesian framework, which are widely used to explain multisensory integration.

Method

Consider an agent with k sensors $O^1,O^2,...,O^k$, where O^i is the observation space of the i^{th} sensor. Furthermore, assume that the environment is fully observable in the Cartesian product of the observation spaces, i.e. $S=O^1\times O^2\times ...\times O^k$. At each time step, the agent should choose an action from its action set A according to the perceptual input (state) $s=(o^1,o^2,...,o^k)$, where o^i is the current reading of the i^{th} sensor. After performing the action, the agent receives an immediate reinforcement signal (reward) r from the environment. It is assumed that all the reward distributions, corresponding to the state-action pairs, are unknown with support in [0,1]. The goal of the agent is to maximize the total amount of reward it receives over its lifetime. To achieve this goal, the agent should learn the appropriate action in response to members of the joint sensory space S.

The primary challenge here is that the state space S is high dimensional. Therefore, to learn the best action corresponding to each member of S, a large number of experiences (samples) is needed. This problem is known as the curse of dimensionality. One way to tackle this problem is to use the experiences in the subspaces of S", such as O^i, for decision making [11], [12]. However, the environment in the eyes of O^i is partially observable, which creates a many-to-one mapping between real states of the environment and observations in O^i. This problem is known as Perceptual Aliasing (PA) [13] and is avoided in general. Nevertheless, PA might be beneficiary in learning a task [11], since it can partially free the learner from the curse of dimensionality if states sharing the same o^i have similar optimal policies. PA might be helpful at the early stages of learning as well, where learning a moderately rewarding policy over O^i is faster than learning a policy with the same reward over the joint space S. In these two cases, learning in the subspaces results in generalization of experiences. In contrast, PA can be very undesirable when functionally different states of the environment, i.e. states with very different policies, are mapped to a same observation in O^i. This case of PA turns the accumulated experience in that subspace into "garbage" [14]. Figure 1 illustrates these concepts in a simple example. Our proposed statistical test (see Generalization Test) has the ability to detect different cases of perceptual aliasing that are illustrated in the figure.

In order to benefit from PA and to avoid its harms, a statistical test is proposed to discriminate estimates of the expected reward which are instances of generalization (beneficial cases of PA) from garbage information. The proposed test is in part inspired from McCallum's work on learning with incomplete perception [15]. Then, a selection policy for choosing the most reliable source of information is employed. Finally, according to the selected information, a decision making policy has been introduced which considers the exploration and exploitation trade-off. A schematic overview of the proposed method, including the Generalization Test (G Test) and the Decision Making phase, is illustrated in Figure 2. In the following subsections, the proposed multisensory learning and decision making method is explained in detail.

In general, there are two approaches for learning a task, learning through labeled samples and learning by interaction. State estimation in a supervised setting requires having the specifications of the states at hand. Nevertheless, in reality we should learn the states either directly or through learning the optimal policy. In the problem at hand, the agent begins its life in a tabula rasa state and there is no information available regarding the observation models of sensors and the relation between the agent's sensory space S and its action space A. Furthermore, the only teacher that the agent can interact with is the environment. Therefore, only through interactions with the environment, the agent can learn to act properly. In this problem we are not interested in learning the observation models of individual sensors nor do we have the necessary sources of feedback to do this. Therefore, this problem is different from the conventional supervised learning where a teacher provides a set of labeled data, and the agent needs only to learn the observation models of sensors and perform a state estimation task.

1. Modeling

The actual value of choosing action $a\in A$ when the agent is in state $s=(o^1,o^2,\ldots,o^k)$ is denoted as $Q^*(a,s\in S)$, and its estimated value as $Q(a,s\in S)$. All the estimated values (Q-values) are represented in a $|O^1|\times|O^2|\times...\times|O^k|\times|A|$ dimensional table, known as Q-table. Q-values are updated after each time step using

$$Q(a,s\in S)=Q(a,s\in S)+\beta(a,s\in S)[r(a,s\in S)-Q(a,s\in S)],$$

where $r(a,s\in S)$ is the reward received after performing a in s, and $0<\beta(a,s\in S)\leq 1$ is the learning rate for the given state and action. We assume that the reward distributions are fixed throughout the learning; i.e. the environment is stationary. In stationary environments, it is rational to employ $\beta(a,s\in S)=\frac{1}{\#(a,s\in S)}$, where $\#(a,s\in S)$ is the sample size, i.e. the number of times that action a is performed in state s. By using this learning rate, the above equation becomes identical to the incremental update formula for computing the average reward [8]. Therefore, Q-values are the sample means and Q^*s are the actual means of the underlying reward distributions.

As it will be explained in the following sections, we need confidence intervals on Q^* s for our generalization test and decision making method. For a moderately large number of samples, we can create a confidence interval on $Q^*(a,s\in S)$ using the following bound [16]:

$$P\left(Q(a,s\in S)-t_{\frac{\alpha}{2}}^{\#(a,s\in S)-1}\times\frac{\overline{std}(a,s\in S)}{\sqrt{\#(a,s\in S)}}\leq Q^*(a,s\in S)\leq Q(a,s\in S)+t_{\frac{\alpha}{2}}^{\#(a,s\in S)-1}\times\frac{\overline{std}(a,s\in S)}{\sqrt{\#(a,s\in S)}}\right)=1-\alpha \quad (1)$$

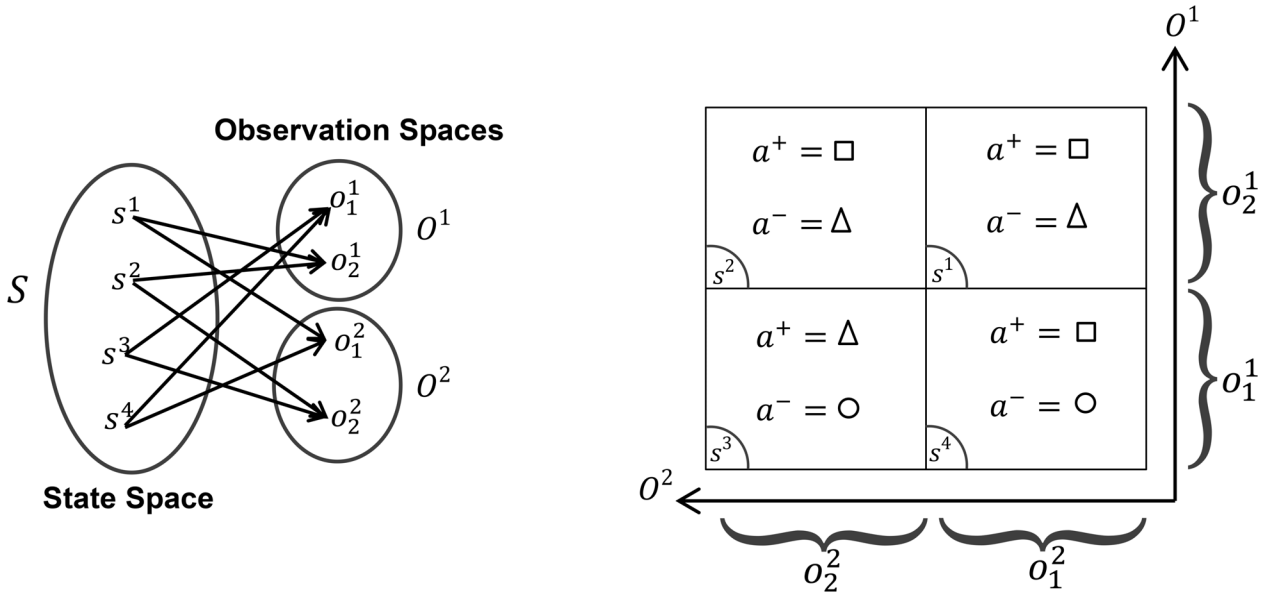

Figure 1. Different types of perceptual aliasing in subspaces. $O^i = \{o_1^i, o_2^i\}$ represents the observation set of the i^{th} sensor for i = 1, 2. $S = \{s^1, s^2, s^3, s^4\}$ is the state set and $A = \{\bigcirc, \square, \Delta\}$ is the action set of the agent. a^+ and a^- are the best and the worst actions in the given state, respectively. Accumulated experience in o_2^1 is a perfect generalization for s^1 and s^2, since these two states have the same optimal policy and o_2^1 is common between them. In contrast, accumulated experience in o_2^2 is garbage information because functionally different states are mapped to the same observation. The situation for o_1^2 and o_1^1 is a little different. Only for the best action in o_1^2 and the worst action in o_1^1 we have the generalization, however, for the other action this is not the case.
doi:10.1371/journal.pone.0103143.g001

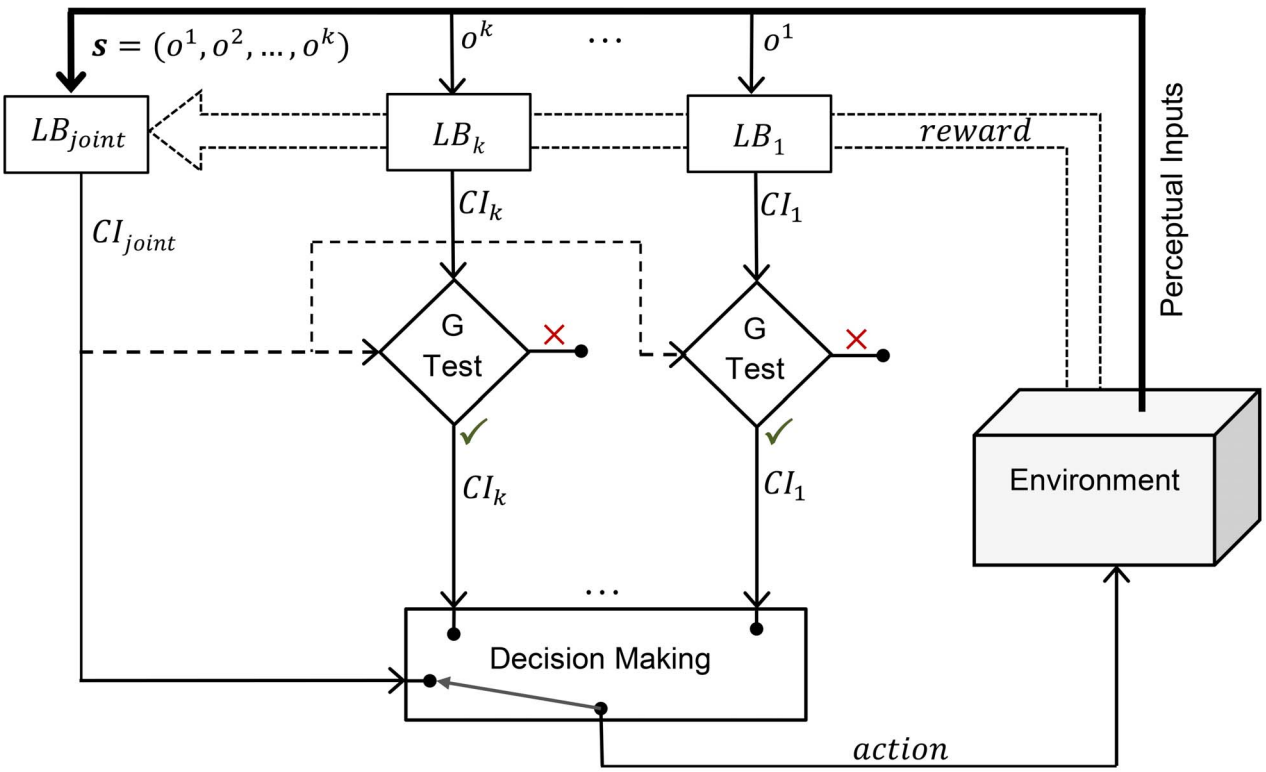

Figure 2. A schematic overview of the proposed framework for multisensory learning and decision making. $s = (o^1, o^2, \ldots, o^k)$ is the perceptual input, o^i is the current reading of the i^{th} sensor, and LB_i is the learning block of the i^{th} sensor. For each action and based on the previously received rewards, each learning block calculates a confidence interval (CI) on the mean of the reward distribution corresponding to the given observation and action pair. The proposed Generalization Test (G Test), tests the generalization ability of the individual source against the joint space. In case that an individual source passes the G Test, its confidence interval will be considered in the decision making phase. In decision making phase, an appropriate action based on the given intervals will be selected which considers the exploration and exploitation trade-off.
doi:10.1371/journal.pone.0103143.g002

In (1) $t_{\frac{\alpha}{2}}^{\#(a,s\in S)-1}$ is the Student t distribution with $\#(a,s\in S)-1$ degrees of freedom. The parameter $\alpha\in[0,1]$ controls the confidence that Q^* will fall inside the confidence interval. Finally, the value $\overline{std}(a,s\in S)$ is the estimated standard deviation of the underlying reward distribution defined by

$$\overline{std}(a,s\in S)=\sqrt{\frac{\#(a,s\in S)\sum r^2(a,s\in S)-(\sum r(a,s\in S))^2}{\#(a,s\in S)\times(\#(a,s\in S)-1)}},$$

where $\sum r(a,s\in S)$ is the sum of the rewards and $\sum r^2(a,s\in S)$ is the sum of the squares of the rewards received by performing a in s.

The confidence interval in (1) is mathematically valid when either the number of samples ($\#(a,s\in S)$) is moderately large or when the reward distribution is Normal (Gaussian). Although these conditions may seem rather restricting, in our experience, bound (1) works reasonably well in most practical cases.

When the sample size is not sufficiently large or the reward distribution is not Gaussian, we may use Chebyshev's inequality to calculate the confidence interval. To do so, we need the true standard deviation of the reward distribution, which is not available in general. However, defining the reward distribution in the interval [0,1], the maximum possible value for the variance is $\frac{1}{4}$. Then a very conservative Chebyshev's inequality is

$$P\left(Q(a,s\in S)-\frac{1}{\sqrt{\alpha}}\times\frac{0.5}{\sqrt{\#(a,s\in S)}}\le Q^*(a,s\in S)\le Q(a,s\in S)+\frac{1}{\sqrt{\alpha}}\times\frac{0.5}{\sqrt{\#(a,s\in S)}}\right)\ge 1-\alpha \quad (2)$$

Although bounds (1) and (2) are similar in essence, bound (2) is very conservative but independent of the reward distribution. Conservativeness of (2) has roots in not taking into account the type of the reward distribution and its estimated variance. This lack of prior assumptions will result in extremely conservative intervals in cases that the variances are very small or even zero. In situations like these, it is better to employ the "variance-aware" inequality proposed in [17]:

$$P\left(Q(a,s\in S)-\overline{std}(a,s\in S)\sqrt{\frac{2\ln\frac{3}{\alpha}}{\#(a,s\in S)}}-\frac{3\ln\frac{3}{\alpha}}{\#(a,s\in S)}\le Q^*(a,s\in S)\le Q(a,s\in S)+\overline{std}(a,s\in S)\sqrt{\frac{2\ln\frac{3}{\alpha}}{\#(a,s\in S)}}+\frac{3\ln\frac{3}{\alpha}}{\#(a,s\in S)}\right)\ge 1-\alpha \quad (3)$$

In this study, we are mainly interested in the *length* of the confidence intervals and their *relative length* to each other. Generally, by visiting new samples, the length of all the intervals in bounds (1), (2), and (3) diminishes gradually. Therefore, as we will see in the following sections, all the mentioned intervals are applicable in our algorithm. In Discussions and Conclusions section, a discussion on a number of practical points concerning these bounds is provided.

For individual sensors, $Q^*(a,o^i\in O^i)$ denotes the actual mean and $Q(a,o^i\in O^i)$ denotes the sample mean of reward, received by performing action $a\in A$ when the i^{th} sensor's observation is o^i. We can create a confidence interval on $Q^*(a,o^i\in O^i)$ by using the same procedure and only replacing the following variables in bounds (1), (2), or (3):

$$\#(a,o^i\in O^i)=\sum_{p^1,\ldots,p^{i-1},p^{i+1},\ldots,p^k}\#(a,p^1\in O^1,\ldots,o^i\in O^i,\ldots,p^k\in O^k) \quad (4)$$

$$Q(a,o^i\in O^i)=\frac{1}{\#(a,o^i\in O^i)}\sum_{p^1,\ldots,p^{i-1},p^{i+1},\ldots,p^k}Q(a,p^1\in O^1,\ldots,o^i\in O^i,\ldots,p^k\in O^k)\#(a,p^1\in O^1,\ldots,o^i\in O^i,\ldots,p^k\in O^k) \quad (5)$$

The above equations express the marginal values for the i^{th} sensor.

In order to calculate $\overline{std}(a,o^i\in O^i)$ we also need to calculate two more terms:

$$\sum r^2(a,o^i\in O^i)=\sum_{p^1,\ldots,p^{i-1},p^{i+1},\ldots,p^k}[\sum r^2(a,p^1\in O^1,\ldots,o^i\in O^i,\ldots,p^k\in O^k)] \quad (6)$$

$$\sum r(a,o^i\in O^i)=\sum_{p^1,\ldots,p^{i-1},p^{i+1},\ldots,p^k}[\sum r(a,p^1\in O^1,\ldots,o^i\in O^i,\ldots,p^k\in O^k)] \quad (7)$$

Calculation of (4)–(7) does not need extra learning trials because, these variables are calculated by marginalization of statistics of the joint space S.

2. Generalization Test

A statistical test is proposed to answer the following question:

Is perceptual aliasing in o^i, a beneficial case of generalization for action $a\in A$,or a harmful case of "garbage" information?

Based on our modeling, we can restate the question as "is $Q^*(a,o^i\in O^i)$ a reasonable representation of $Q^*(a,s\in S)$?", where o^i is the current observation of the i^{th} sensor and $s=(o^1,o^2,\ldots,o^k)$. However, as previously mentioned, Q^* s are unknown. As such, we use their confidence intervals by employing either bounds (1), (2), or (3). We denote the confidence interval on $Q^*(a,s\in S)$ as M and confidence interval on $Q^*(a,o^i\in O^i)$ as CI_i.

To validate the generalization ability of CI_i, we need to test whether CI_i and M are estimating the same value ($Q^*(a,s\in S)$). However, due to perceptual aliasing (many-to-one mapping), CI_i has also experienced all the rewards used in the calculation of M. Hence, checking the significance of their difference does not provide useful information. The proposed idea here is to extract the common experiences between CI_i and M, and then perform a statistical test on the residuals of CI_i, and M. The procedure of extracting common experiences from CI_i is as follows:

$$Q'(a,o^i\in O^i)=\frac{\#(a,o^i\in O^i)Q(a,o^i\in O^i)-\#(a,s\in S)Q(a,s\in S)}{\#(a,o^i\in O^i)-\#(a,s\in S)} \quad (8)$$

Table 1. The function that implements MOS method.

function MOS(*M, Accepted*)

Input: $_M$ is the confidence interval on the joint space, *Accepted* is the array storing confidence intervals on the sources that passed the generalization test

1: $MOS \leftarrow \arg \max_{CI \in Accepted} \overline{CI}$

2: $v \leftarrow \min(\overline{MOS}, \overline{M})$

3: return v

doi:10.1371/journal.pone.0103143.t001

$$\#'(a,o^i \in O^i) = \#(a,o^i \in O^i) - \#(a,s \in S) \quad (9)$$

$$\sum r^{2'}(a,o^i \in O^i) = \sum r^2(a,o^i \in O^i) - \sum r^2(a,s \in S) \quad (10)$$

$$\sum r'(a,o^i \in O^i) = \sum r(a,o^i \in O^i) - \sum r(a,s \in S) \quad (11)$$

By using the variables on the left side of the above equations, a new confidence interval CI_i' can be created using any of bounds (1), (2), or (3). For each action, CI_i' represents the intervallic estimate of the mean of a reward distribution created from experiences in the current observation of the i^{th} sensor, minus the experiences in the current state of the environment. If there exists an intersection between CI_i' and M, then there is a good chance that CI_i and M are estimating the similar expected value of rewards ($Q^*(a,s \in S)$). In other words, it means that the perceptual aliasing in CI_i is a case of generalization. The proposed test states that at each time step for action a:

$$\text{Reject} CI_i \Leftrightarrow M \cap CI_i' = \varnothing \quad (12)$$

Based on (12), we can expect the following behavior in different stages of learning:

- During initial steps of learning (when sample size is very small), M and CI_i' both have large confidence intervals. Consequently, CI_i will be able to pass the proposed test in most time steps. Due to the low uncertainty in CI_i, this behavior is desirable during initial steps.
- By gaining new samples, both M and CI_i' shrink. Therefore, the i^{th} sensor will be able to pass the test only if its experiences are a good generalization of M's experience.
- As the sample size for M increases, its interval becomes smaller and smaller to a degree where it dwindles to only contain $Q^*(a,s \in S)$. The same thing happens for CI_i' but it will converge to a different point. As a result, the test will reject all the individual sensors.

3. Decision Policy

As mentioned earlier, the agent starts with no prior information about the environment and the task at hand. Consequently, throughout the learning it faces the dilemma of gaining new experience by choosing one of the less explored decisions or exploiting the past experiences by selecting one of the well-rewarded decisions. This problem is known as the exploration versus exploitation trade-off [8].

At each state $s \in S$, it can be assumed that there are $|A|$ unknown reward distributions which correspond to each action in the action set A. The best action a^* is the one corresponding to the distribution with the greatest mean, i.e. $a^* = \arg\max_{a \in A} Q^*(a,s \in S)$. However, Q^* s are unknown and the agent should make the decision based on their estimates. A good decision policy should consider both the Q-value (sample mean statistic) and the uncertainty regarding its expected value. The value of the sample mean controls the exploitative selections, while its uncertainty controls the explorative decisions. Clearly, the uncertainty of the sample mean tends to zero as the number of samples tends to infinity, resulting in a smooth transition from exploration to exploitation as the number of samples increases.

A well-studied family of decision policies, which considers these two criteria, works based on the idea of creating an upper confidence interval on the mean of each reward distribution. Based on the calculated upper bounds, the decision policy selects the action with the greatest upper confidence interval [18]. This

Table 2. The function that implements LUS method.

function LUS(*M, Accepted*)

Input: $_M$ is the confidence interval on the joint space, *Accepted* is the array storing confidence intervals on the sources that passed the generalization test

1: $LUS \leftarrow \arg \min_{CI \in Accepted} (\overline{CI} - \underline{CI})$

2: **if** $(\overline{M} - \underline{M} < \overline{LUS} - \underline{LUS})$ **then** $LUS \leftarrow M$

3: $v \leftarrow \min(\overline{LUS}, \overline{M})$

4: **return** v

doi:10.1371/journal.pone.0103143.t002

Table 3. The proposed Algorithm for Multisensory Learning and Decision Making.

Initialize $Q(a,s)$, $\#(a,s)$, $\sum r(a,s)$, and $\sum r^2(a,s)$ to zero $\forall s\in S, a\in A$	
1:	**Repeat** at each time step
2:	$s=(o^1,o^2,\ldots,o^k)$
3:	**for each** $a\in A$ **do**
4:	$Accepted\leftarrow\varnothing$
5:	**for each** sensor i **do**
6:	Calculate M, CI_i, and $CI_i^{'}$ based on either bounds (1), (2), or (3)
7:	**if** $(M\cap CI_i^{'})\neq\varnothing$ **then** $Accepted\leftarrow Accepted\cup\{CI_i\}$
8:	$value(a)\leftarrow$ MOS$(M,Accepted)$ or LUS$(M,Accepted)$
9:	Perform $a^+=\arg\max_{a'\in A} value(a')$, observe reward r
10:	$\#(a^+,s)=\#(a^+,s)+1$
11:	$\sum r(a^+,s)=\sum r(a^+,s)+r$
12:	$\sum r^2(a^+,s)=\sum r^2(a^+,s)+r^2$
13:	$Q(a^+,s)=\sum r(a^+,s)\big/\#(a^+,s)$
14:	**Until** the end of the learning

doi:10.1371/journal.pone.0103143.t003

idea is known as "optimism in face of uncertainty principle." It has been proved that variations of these decision policies, such as UCB1 [19], achieve logarithmic expected regret, i.e. the expected loss due to the fact that the agent does not always choose the optimal action, uniformly over the total number of samples of the given state. This amount of regret is the smallest possible expected regret, up to a constant factor. Fortunately, in the proposed approach we have already employed confidence intervals on the means of the reward distributions. The only difference in our problem is that we have a set of confidence intervals, instead of one, for each action. Therefore, we need to integrate available confidence intervals to one, and then employ the mentioned idea.

One can devise various methods for integrating a set of intervals. However, in this study we are interested in finding, specifically, the source of information that has the greatest impact on the final decision. As a result, we reduce the integration problem to selection of one of the available intervals as the representative interval for the given action. We propose two methods for this interval selection. The first method works by the idea of selecting the Most Optimistic Source (MOS), while the second method chooses the Least Uncertain Source (LUS). Details of these methods are as follows:

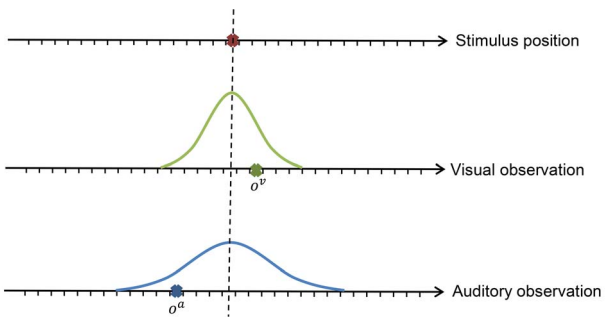

Figure 3. Stimulus and observations by the auditory (o^a) and the visual (o^v) sensors. Observations are based on Gaussian noise models. Variances control the reliability of each sensor.
doi:10.1371/journal.pone.0103143.g003

At each state $s\in S$ and for each action $a\in A$, given a set of confidence intervals of individual sensors which were able to pass the previously mentioned test (12), the MOS method selects the interval with the greatest upper bound. The LUS method, on the other hand, selects the interval with the shortest length. The upper bound value of the selected interval will be used as the representative value for action a. However, if this value is greater than M's upper bound, then M's upper bound will be used as the representative value. The reason behind this constraint is that, regardless of its great uncertainty, M is still the most reliable (with lowest aliasing) source of information regarding the actual mean of the underlying reward distribution. Therefore, any value greater than M's upper bound is unrealistically optimistic. The idea behind LUS is that shorter intervals indicate lower uncertainty, and it is always desirable to attend the least uncertain source of information for decision making. The pseudo-codes of the MOS and LUS methods are shown in Table 1 and Table 2. For bound B, the notations $\overline{B}$ and $\underline{B}$ represent the upper bound and lower bound values of B, respectively.

After choosing an upper bound value (with either MOS or LUS methods) for all the actions, the action with the maximum upper bound value is selected as the final decision. By performing the selected action, the environment returns the reward $r\in[0,1]$. The complete pseudo-code of the proposed method is shown in Table 3. The only parameter that needs to be initialized is $\alpha\in[0,1]$, where $1-\alpha$ is the confidence coefficient of confidence intervals.

Experiments and Results

The task is a modified version of the localization task in the visual and auditory modalities [2] [20]. The simulation setup is based partly on [10]. At each time step, a stimulus is generated randomly in one of the 30 discrete positions and each sensor observes a noisy representation of it. The observation noise for each sensor is modeled by a Gaussian distribution with standard deviation δ; see Figure 3. After observing the stimulus through its sensors, the agent chooses one of the 30 discrete positions as the

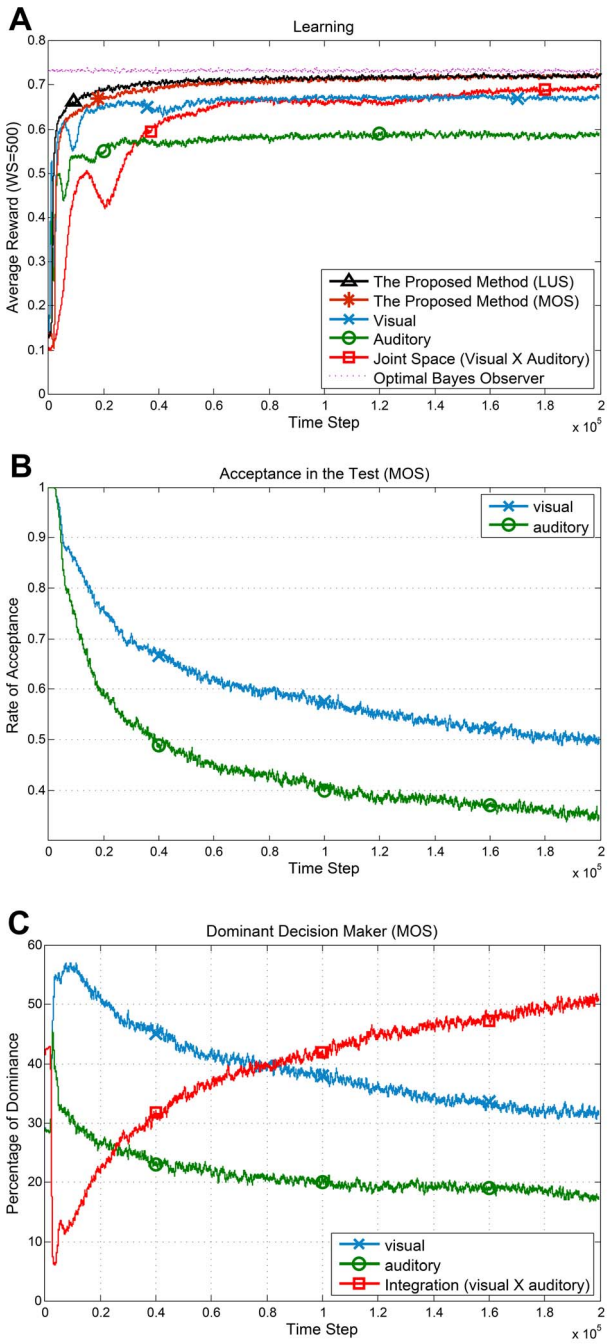

Figure 4. Performance and behavior of the method in the localization task. All graphs are results of averaging over 20 independent runs and passing a moving average window with size 500. (**A**) Average reward for all agents. For the proposed methods (MOS and LUS), we used Table 3, employing bound (1) with $\alpha=0.1$ for calculating confidence intervals. The rival methods employ the UCB1 policy on the individual sensors and on the joint space. (**B**) Average acceptance rate (1–rejection rate) of the individual sensors in the proposed method (MOS). (**C**) The average dominancy percentage of each source in decision making (MOS). In the first half of learning steps, vision is the dominant sensor while the agent prefers the integrated sensory data in the rest of learning steps.
doi:10.1371/journal.pone.0103143.g004

desirable action and receives an immediate reinforcement value in $[0,1]$:

$$reward = \max(0,(1-\frac{1}{\tau}\times|\text{action}-\text{stimulus position}|)) \qquad (13)$$

We used $\tau=4$, which indicates that only actions (estimates) within a radius of three units from the stimulus position receive positive rewards.

The agent has no prior information about the task, the observation models, and the relation between the sensory space and actions. Therefore, throughout the learning, it should learn the appropriate action only based on the sensory inputs and previously received rewards. On the other hand, the optimal Bayesian observer [2] assumes that all of the mentioned information is available and chooses its action according to the following integration rule:

$$action = \frac{1/\delta_a^2}{1/\delta_a^2+1/\delta_v^2}o^a + \frac{1/\delta_v^2}{1/\delta_a^2+1/\delta_v^2}o^v, \qquad (14)$$

where δ_a and δ_v are the standard deviations of the Gaussian noise models for the auditory and visual inputs, respectively. Moreover, o^a and o^v are the representations of the stimulus in the auditory and visual observation spaces. Behavioral studies have shown that adults integrate information from sensors in a statistically optimal manner which based on the Gaussian observation models, can be formulated by equation (14).

In all the following experiments, the proposed method uses the Cartesian product of the observation spaces of all the sensors for its state space. The agent's learning and decision making is based on Table 3.

Experiment 1

In the first experiment we use $\delta_v^2=3$ and $\delta_a^2=5$ (see Figure 3). In order to validate our method, we employ three different agents. Two of the agents (Visual and Auditory agents) use only the individual sensors which will result in a state-action space of size 30×30 for each. The third one (Visual $\times$ Auditory agent) uses both sensors for its learning and decision makings and has a state-action space of size $30\times30\times30$. For these three agents, we employ the UCB1 policy [19] for decision making. UCB1 calculates upper bounds on the means of the reward distributions based on the Hoeffding inequality. At each state s, UCB1 chooses the action that maximizes

$$upperBound(a) = Q(a,s\in S) + \sqrt{\frac{\rho\times\ln(\sum_{a'\in A}\#(a',s\in S))}{\#(a,s\in S)}}, \qquad (15)$$

where $Q(a,s\in S)$ is the average reward obtained from performing action a in state s, $\#(a,s\in S)$ is the number of times a has been selected in s, and ρ is the exploration coefficient [17]. In the original version of UCB1, ρ is set to 2. However, this value results in a high exploration rate. We use $\rho=0.2$ in all the experiments to increase the speed of learning for the rival agents.

It should be noted that when we use initial capital for a sensor, we are referring to the agent that learns in that sensor space. For instance, Visual refers to the agent that uses only the visual sate space for its learning.

Table 4. Analyzing the learning speed and the behavior of different methods for Experiment 1 and 2.

Percentage of accumulated reward	Learning Method	Experiment 1				Experiment 2				
		# time step	Percentage of dominance			# time step	Percentage of dominance			
			V	A	I		V	A	N	I
60%	Joint Space	38,113	0	0	100	1,141,640	0	0	0	100
	MOS, bound(1), $\alpha=0.1$	8,200	56	32	12	12,455	62	27	7	4
	LUS, bound (1), $\alpha=0.1$	5,010	56	32	12	5,557	61	32	5	2
	MOS, bound (2), $\alpha=0.4$	7,901	62	37	1	10,828	60	32	8	0
	LUS, bound (2), $\alpha=0.4$	5,599	64	35	1	8,852	60	29	11	0
75%	Joint Space	81,179	0	0	100	2,437,811	0	0	0	100
	MOS, bound (1), $\alpha=0.1$	17,393	52	28	20	33,911	62	25	4	9
	LUS, bound (1), $\alpha=0.1$	10,341	58	29	13	17,289	57	31	5	7
	MOS, bound (2), $\alpha=0.4$	17,854	61	37	2	35,979	67	28	4	1
	LUS, bound (2), $\alpha=0.4$	14,138	67	31	2	35,461	68	27	4	1
90%	Joint Space	348,945	0	0	100	10,036,225	0	0	0	100
	MOS, bound (1), $\alpha=0.1$	72,689	40	20	40	1,148,066	43	20	2	35
	LUS, bound (1), $\alpha=0.1$	43,281	50	25	25	974,986	39	21	4	36
	MOS, bound (2), $\alpha=0.4$	96,437	53	38	9	1,767,754	58	30	3	9
	LUS, bound (2), $\alpha=0.4$	94,204	66	25	9	1,831,145	61	24	3	12

The performance criterion is the number of time steps needed to reach a certain percentage of the Bayesian optimal observer's accumulated reward. V = visual, A = auditory, N = noise, I = integration.

doi:10.1371/journal.pone.0103143.t004

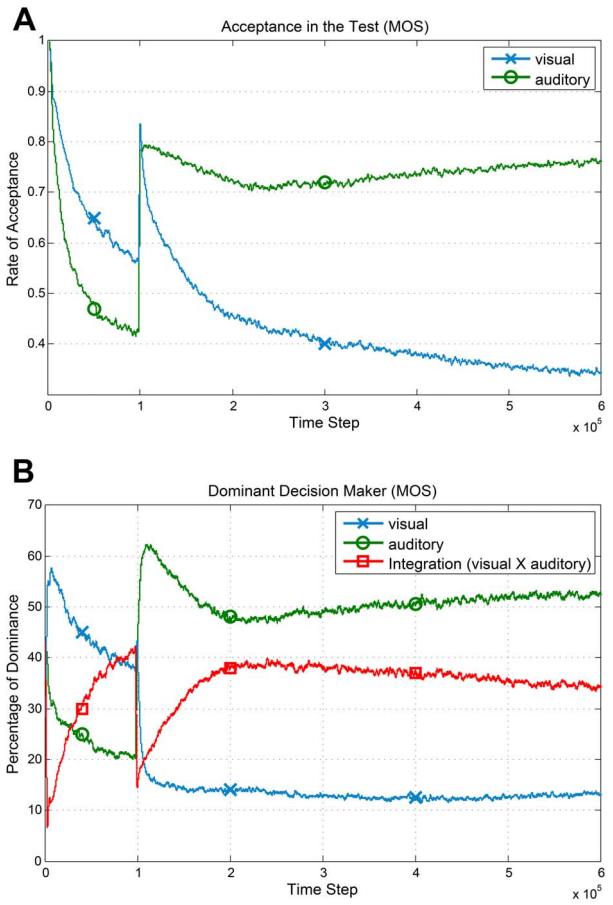

Figure 5. Performance of the method (MOS) in response to an unexpected change in the environment. At time step 10^5 the visual sensor fails and its variance changes to the highest possible value. All graphs are results of averaging over 10 independent runs and passing a moving average window with size 500. (**A**) Average acceptance rate ($1-$rejectionrate) of the individual sensors. (**B**) The average dominancy percentage of each source in decision making (MOS). After failure of the visual sensor, the method detects this change and relies on the auditory sensor for decision making.
doi:10.1371/journal.pone.0103143.g005

The average reward against the time step for all the agents and the optimal Bayesian observer are shown in Figure 4A. For the proposed methods (MOS and LUS), we employed bound (1) with $\alpha=0.1$. As can be seen in the figure, the proposed methods have a noticeably faster learning and higher rewards compared to the Visual $\times$ Auditory agent. The Visual and the Auditory agents both have a smaller state space (only one sensor) which results in a fast learning during initial time steps. However, due to their partial perception, they can never reach the performance of the optimal Bayesian observer.

To evaluate the proposed generalization test (see Figure 2 and Generalization Test) for the proposed method (MOS), the average outcome of the test for the chosen action against the time step is shown in Figure 4B. The value in the vertical axis specifies the rate of acceptance in the test which is 1–rejection rate. The test completely accepts the individual sensors during initial steps. This is in line with having a generalization power in the individual sensors due to more samples. Nevertheless, as the joint space learning improves, the rate of acceptance for the individual sensors decreases. This is because of sufficient experience accumulation in the joint space and existence of perceptual aliasing in the individual sensor spaces. This decline is more noticeable for the auditory sensor which is less reliable.

To investigate the decision making behavior of the proposed method (MOS), the average dominancy percentage of each source of information over time is shown in Figure 4C. In the initial steps of learning, vision is the dominant modality. However, as the time step increases there is a tendency to rely on the joint space for decision making (sensory integration). Considering Figure 4A and Figure 4C we can conclude that as the average reward received in the joint space increases, the proposed method gradually switches its decision policy from selection to integration. This behavior is comparable to the humans' shift from sensory selection at childhood to sensory integration at adulthood.

Performance criteria for different variations of the proposed method and the Visual $\times$ Auditory agent are illustrated in Table 4.

In Figure 4A there is a temporary decline in the average reward of the individual sensors and the joint space agents. The reason behind these declines is the inherent temporary exploration in UCB1. In UCB1, the policy calculates $1-\alpha$ upper confidence bound where α has an inverse relation with the total number of samples in state s (the logarithmic term in equation (15)). Therefore, if an action has not been visited in a state for a long time, this term forces the agent to choose that action. For large state-action spaces, it creates temporary exploration phases in the learning. This exploration is beneficial in non-stationary environments, however, our environment is stationary and the exploration results in the observed decline. We reduced the exploration effect by using small ρ in (15). We tested the individual sensors and the joint space agents using constant alpha and different types of confidence intervals as well and the significant superiority of the proposed method was still intact.

A non-stationary change in the environment. Having a stationary environment is one of the basic assumptions we made. To investigate the effect of an unexpected change in the environment, we decreased the reliability of visual sensor to the lowest possible value at step 10^5. The underlying reward distributions for the visual sensor and the joint space changed accordingly. As Figure 5A shows, this change is detected by the proposed test. As a result, the rate of acceptance of the visual sensor noticeably decreases after step 10^5. However, in the decision making section, only the MOS method could cope with this disturbance and the LUS method failed to adapt its behavior; as it relies more on the joint space. The percentage of dominance for each source of information in the MOS method is shown in Figure 5B. After time step 10^5, the agent relies more on the auditory sensor and only about 13% of decisions are made according to the visual data. We will discuss more on non-stationary environments in Discussions and Conclusions.

Parameter setting. The method (Table 3) does not need any tuning and the only open parameter is $\alpha\in[0,1]$, initialized at the beginning of the learning. Alpha defines the agent's characteristic; smaller value for α results in larger confidence intervals which means more tendency toward exploration than exploitation. Moreover, small value for alpha makes the test easier for individual sensors to pass, and as a result, postpones the transition from selection to integration. Figure 6 shows these effects in Experiment 1.

Experiment 2

The goal of this experiment is to study the method in the presence of an added unreliable sensor (noise). The new sensor's reading is uniformly distributed noise. In other words, there is no correlation between the position of the stimulus and the sensor's

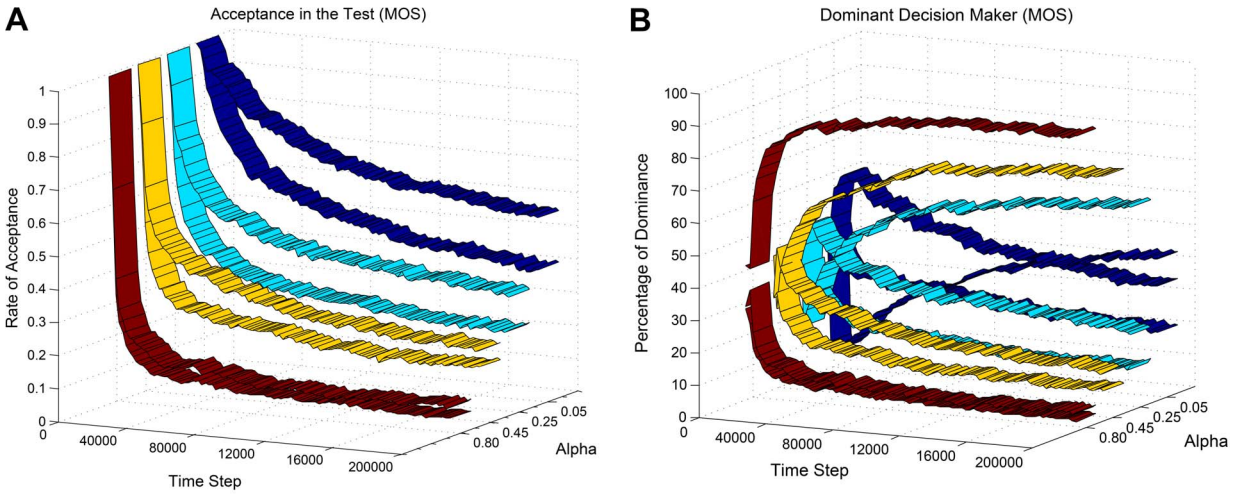

Figure 6. Impact of α. We used four different values (0.05, 0.25, 0.45, 0.80) for α from being conservative to liberal in terms of confidence. All graphs are results of averaging over 10 independent runs and passing a moving average window with size 500. (**A**) Average acceptance rate (1–rejection rate) of the individual sensors in the proposed method (MOS). The upper/lower ribbon for each value of α represents visual/auditory sensor. By increasing α, the test becomes harder for the individual sensors to pass. (**B**) The average dominancy percentage of each source in decision making (MOS). For each value of α, the ascending ribbon represents integration and the two descending ribbons represent selection of visual and auditory sensors. Increasing α results in earlier cross of the ascending and the descending ribbons; i.e. earlier switch from selection to integration.
doi:10.1371/journal.pone.0103143.g006

reading. By adding this sensor, the size of the joint state-action space jumps to $30\times30\times30\times30$.

The Noise agent has no beneficial learning and its average reward curve is flat throughout its life; see Figure 7A. Furthermore, due to the presence of this unreliable sensor, learning by the joint space agent has been drastically diminished compared to the Visual agent. The proposed method (MOS) has been able to identify the unreliable source of information and therefore, has been superior to the joint space agent in terms of both learning speed and average reward. However, during the initial steps of learning, its average reward is slightly lower than the Visual agent. It is the cost of having no prior information about the unreliable sensor which makes the method to explore more at the early steps of learning.

The results of the proposed test and the percentage of dominance of each source of information in decision making are shown in Figure 7B and Figure 7C, respectively. The rate of acceptance for all subspaces declines by time and this decline is faster for the unreliable sensor. Moreover, according to Figure 7C, only about 3% of the time the unreliable sensor chooses the final decision. This noise selection mostly contains explorative decisions. This result is evidence that the proposed method clearly considers a subsection of its state space as unreliable and filters it in the decision makings.

Comparisons. Table 4 illustrates learning speed in terms of the number of time steps required for each method to reach a certain percentage of the accumulated reward that the Bayesian optimal decision maker achieves. Table 4 also shows the percentage of dominance for each source of information. In all variations of the proposed method, the percentage of dominance for sensory integration increases by progress of learning. Also in the second experiment, the dominance of the noise sensor decreases with time steps. The results indicate that presence of the unreliable sensor in the joint space has made the method slower in the second experiment. This is because the agent has to live with its reliable individual sensors until its joint space yields a reasonable amount of samples to be considered reliable.

We proposed two methods for decision making; namely MOS and LUS, see Table 1 and Table 2. The MOS method chooses the most optimistic source of information, while LUS attends the source with the lowest uncertainty. Both of these criteria are plausible choices for decision making and in our experience both and even some combinations of them work well in practice. Based on Table 4, the LUS method requires fewer time steps compared to the MOS method to reach a certain percentage of performance in both experiments.

Confidence intervals. Due to the extreme conservative nature of bounds (2) and (3), for the same α, their learning speed is slower than bound (1) in most cases. On the bright side, these bounds are mathematically valid for all kinds of reward distributions. To compensate for this conservativeness, it is recommended to use larger values for α (smaller confidence coefficients) when employing bounds (2) and (3). Furthermore, as mentioned in Method Section, bound (3) is only appropriate in situations where the variances of the reward distributions are small. However, in most cases, there is no information available about the type of the reward distributions and their variances. In these general situations, bound (2) with a moderate value for α is a reasonable choice. For example, in both of the discussed experiments, by using bound (2) and increasing the value of α to 0.4, we achieved similar learning speed and average reward to those illustrated in Figure 4A and Figure 7A. A summary of these results is shown in Table 4.

Extension to the power set of sensors. Throughout this paper, only individual sensors along with their joint space were considered as the sources of information. However, by a slight modification in equations (4)–(7), we can calculate the necessary marginal values for any combination of sensors. Based on this idea, instead of k sensors, we can create 2^k-2 sources of information beside the primary joint space. By employing these sources instead of the individual sensors in line 5 of Table 3, a new variation of the proposed method will be formed. Considering this modification in the algorithm, we performed Experiment 2 with the LUS method using bound (1) and $\alpha=0.1$. The percentage of dominance of each source of information is shown in Figure 8. In the first section of

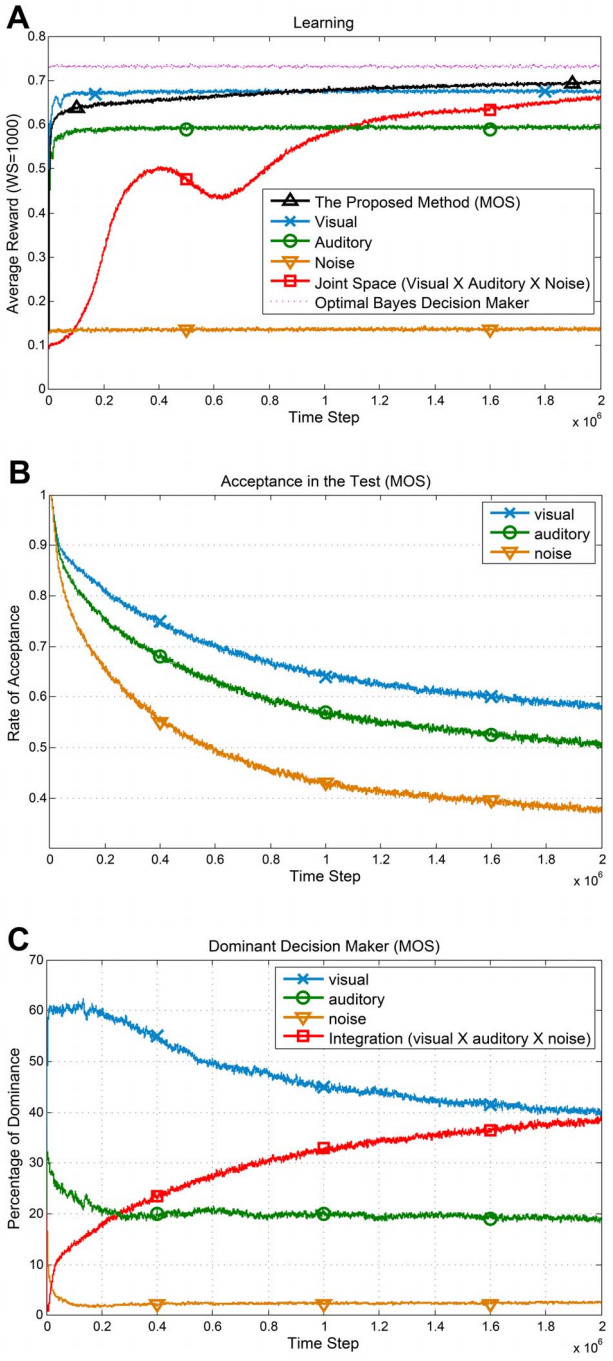

Figure 7. Performance and behavior of the method in response to an unreliable sensor. All graphs are results of averaging over 20 independent runs and passing a moving average window with size 1000. (**A**) Average reward for all agents. For the proposed method (MOS), we used Table 3, employing bound (1) with $\alpha = 0.1$ for calculating confidence intervals. The rival methods employ the UCB1 policy on the individual sensors and on the joint space. (**B**) Average acceptance rate (1–rejection rate) of the individual sensors in the proposed method (MOS). (**C**) The average dominancy percentage of each source in decision making (MOS). Due to unreliability of the noise sensor, it takes longer for learning in the integrated states to mature and, therefore, dominancy of the visual sensor is prolonged.
doi:10.1371/journal.pone.0103143.g007

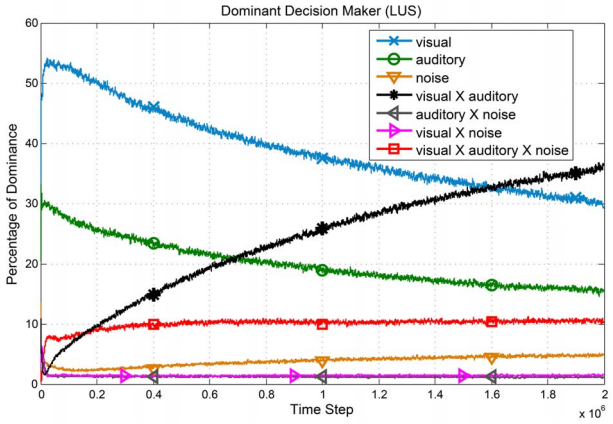

Figure 8. Dominancy of subspaces over time. The average dominancy percentage of different combination of sensors in decision making (LUS). Subspaces including the unreliable source have been filtered. Furthermore, dependency on the integration of reliable sensors increases over time.
doi:10.1371/journal.pone.0103143.g008

learning, the final decision is mostly based on the reliable individual sensors and vision is the dominant modality. However, as the agent matures, the most reliable source of information, which is visual × auditory subspace, takes the main role in decision makings. It means that the extended method has the ability to autonomously elicit the reliable subspaces and to filter the unreliable subspaces of its state space. This modification does not change the amount of required memory. However, the new processing complexity will be exponential, which is still reasonable for tasks with a few sensors.

Discussions and Conclusions

The optimal multisensory integration behavior of adults has been substantially addressed in the literature [1], [2]. However, there are fewer studies and experiments regarding the idea of sensory selection in children [3]–[6]. This lack of sufficient observations is even more significant in the complete age spectral. As a result, there is not sufficient experimental data available to form a definite hypothesis about the transition from sensory selection to sensory integration.

One hypothesis regarding this transition has been proposed by Gori et al. [4], [21]. Their hypothesis is that children select the more accurate sense in multisensory tasks with the purpose of cross-sensory calibration between senses. They suggested that the cross-sensory calibration might have an important impact on maturation of the multisensory perception. In this paper, we have illustrated that even in absence of the cross-sensory calibration hypothesis, the mere transition from the accurate subspaces to the joint space has its own computational advantages. This smooth transition not only facilitates maturation of the multisensory perception, but it is also essential for having a rewarding life.

To show these advantages, we proposed a general multisensory learning method (see Method and Table 3). The proposed method has the ability to autonomously choose different subsets of its state space based on their generalization property and reliability for decision making. Unlike the Bayesian framework, our method neither makes any prior assumptions about the observation model of sensors nor about the relation between sensory space and actions.

It was shown that for an agent who starts its life in a tabula rasa state, the seemingly optimal behavior is to rely on its individual

sensors during early life, and to switch to the joint space (sensory integration) in later stages. This behavior is compatible to the empirical findings. Experimental data indicate that children do not integrate sensory information and make their judgments based only on one sensor, whereas adults use multisensory integration for their decision making [3]–[6]. It was also shown that the proposed method is significantly superior to the individual sensor agents (sensory selection alone) and the joint space agent (only sensory integration) in terms of both learning speed and average reward. Based on these findings, we suggest that this selection and integration, which may be interpreted as two separate methods for decision making, are in fact two sides of a coin and both serve the reward maximization behavior. In addition, the transition from selection to integration is a developmental phenomenon and is smooth.

In our framework, the integration-based decisions will become dominant only after the agent receives enough multisensory experiences during the initial stages of its life. There is also similar empirical evidence that the maturation of the integration decisions is related to the early life experiences (see [22], [3]). Moreover, in [10] the authors showed that by using the reward dependent framework, the problem of causal inference in multisensory perception [23] could also be solved in an interactive fashion. For showing this, they used an artificial neural network for calculating the average reward statistics in the joint sensory space. Based on the average rewards, they used a softmax policy for decision making. With some simplifications, we can say that their agent is inherently equivalent to the joint space agent used in our work. The main focus of Weisswange et al. [10] is on the ability of the learning agent to reach the performance of the Bayesian optimal observer. In our work, on the other hand, we have investigated the role of subspace selection in efficiency of interactive learning. Our results justify that our method can reach the performance of the Bayesian optimal observer as well. On top of that, our method justifies the switch from selection to integration in terms of reward maximization. These studies along with our results indicate that by considering the reward dependent framework, we can model (at least in the behavioral aspect) most of the age-related sensory integration phenomena, without making unnecessary mathematical assumptions about the sensor system and the task.

In Experiment 2 it was shown that the algorithm is also plausible in situations where there is a completely unreliable source of information in the joint space. Even in this extreme scenario, our method outperforms its competitors but faces a slight decrease in the learning speed during initial steps. This decrease is indispensable for any interactive learning method which explores different sources of information.

We assumed that the environment is stationary; i.e. the reward distributions are time invariant, or in other words, the sensory models are fixed throughout the learning. These assumptions are widely used in the learning literature. Nevertheless, interactive learning methods can inherently track non-stationary situations; but of course with a lag due to being experienced-based. We discuss this point more in the sequel. In Figure 5 it is shown that the algorithm (using MOS) tracks the sudden change in the environment, called unexpected uncertainty [24], and adapts itself. Nevertheless, there are some methods to directly deal with unexpected uncertainty. For example, a solution is recalculation of the required statistics after detection of an unusual behavior from the environment. This can easily be done by saving the received rewards in a moving window (a short-term memory) and calculating the necessary statistics accordingly [25].

In this work, for simplicity, we used tables for storing the required statistics. This naturally results in the discretization of the state space. Nevertheless, our approach can be generalized to continuous spaces by using the idea of function approximation for estimating the required statistics in Table 3. We believe that to demonstrate the subspace selection behavior of the proposed method for the task at hand, a simple discrete state space is a well-suited balance of complexity and simplicity. However, in our future works we will investigate and test the theory of continuous version of our algorithm in more complex and practical tasks.

In summary, the proposed algorithm is a dynamic subspace selection method for decision making in interactive learning frameworks. Our method intelligently evades the curse of dimensionality problem by exploiting inherent perceptual aliasing in subspaces. This results in fast learning in addition to an efficient and self-governing transition from sensory selection to integration. This transition is essential for having a rewarding life. In addition, the proposed algorithm (Table 3) is easily implementable. These properties make our method an appropriate candidate for lifetime learning of artificial agents having a large number of sensors. Therefore, an important direction of our research team is to extend the current single-step algorithm to a general multi-step learning and decision making algorithm (reinforcement learning). Based on the value-based decision making framework proposed in [9], we can categorize the main contribution of our algorithm in the representation phase where given a set of sensory inputs, the goal is to achieve the most rewarding state representation.

Acknowledgments

The author Pedram Daee would like to thank Amin Niazi and Habib Zafarian for their time and comments.

Author Contributions

Conceived and designed the experiments: PD MNA. Performed the experiments: PD. Analyzed the data: PD MNA MSM. Contributed to the writing of the manuscript: PD MNA MSM. Developed the model: PD MNA MSM.

References

1. Ernst MO, Banks MS (2002) Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415: 429–433.
2. Alais D, Burr D (2004) The Ventriloquist Effect Results from Near-Optimal Bimodal Integration. Current Biology 14: 257–262.
3. Burr D, Gori M (2012) Multisensory Integration Develops Late in Humans. In: Murray MM, Wallace MT, editors. The Neural Bases of Multisensory Processes. Boca Raton (FL): CRC Press.
4. Gori M, Del Viva M, Sandini G, Burr DC (2008) Young Children Do Not Integrate Visual and Haptic Form Information. Current Biology 18: 694–698.
5. Nardini M, Jones P, Bedford R, Braddick O (2008) Development of Cue Integration in Human Navigation. Current Biology 18: 689–693.
6. Nardini M, Bedford R, Mareschal D (2010) Fusion of visual cues is not mandatory in children. PNAS 107: 17041–17046.
7. Ernst MO (2008) Multisensory integration: a late bloomer. Current Biology 18: R519–521.
8. Sutton RS, Barto AG (1998) Reinforcement Learning: An Introduction. Cambridge, UK: MIT Press.
9. Rangel A, Camerer C, Montague PR (2008) A framework for studying the neurobiology of value-based decision making. Nature Reviews Neuroscience 9: 545–556.
10. Weisswange TH, Rothkopf CA, Rodemann T, Triesch J (2011) Bayesian Cue Integration as a Developmental Outcome of Reward Mediated Learning. PLoS ONE 6(7): e21575. doi:10.1371/journal.pone.0021575
11. Firouzi H, Ahmadabadi MN, Araabi BN, Amizadeh S, Mirian MS, et al. (2012) Interactive Learning in Continuous Multimodal Space: A Bayesian Approach to Action-Based Soft Partitioning and Learning. Autonomous Mental Development, IEEE Transactions on 4: 124–138.

12. Mirian MS, Ahmadabadi MN, Araabi BN, Siegwart RR (2010) Learning Active Fusion of Multiple Experts' Decisions: An Attention-Based Approach. Neural Computation 23: 558–591.
13. Whitehead SD, Ballard DH (1991) Learning to Perceive and Act by Trial and Error. Machine Learning 7: 45–83.
14. Mccallum RA (1995) Instance-Based Utile Distinctions for Reinforcement Learning with Hidden State. In Proceedings of the Twelfth International Conference on Machine Learning: 387–395.
15. Mccallum RA (1993) Overcoming Incomplete Perception with Utile Distinction Memory. In Proceedings of the Tenth International Conference on Machine Learning: 190–196.
16. Casella G, Berger RL (1990) Statistical inference. Belmont, CA: Duxbury Press.
17. Audibert J-Y, Munos R, Szepesvári C (2009) Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. Theoretical Computer Science 410: 1876–1902.
18. Lai T, Robbins H (1985) Asymptotically efficient adaptive allocation rules. Advances in Applied Mathematics 6: 4–22.
19. Auer P, Cesa-Bianchi N, Fischer P (2002) Finite-time Analysis of the Multiarmed Bandit Problem. Machine Learning 47: 235–256.
20. Battaglia PW, Jacobs RA, Aslin RN (2003) Bayesian integration of visual and auditory signals for spatial localization. Journal of the Optical Society of America A, Optics, image science, and vision 20: 1391–1397.
21. Gori M, Sandini G, Burr D (2012) Development of Visuo-Auditory Integration in Space and Time. Frontiers in Integrative Neuroscience 6: 77.
22. Wallace MT, Stein BE (2007) Early experience determines how the senses will interact. Journal of Neurophysiology 97: 921–926.
23. Körding KP, Beierholm U, Ma WJ, Quartz S, Tenenbaum JB, et al. (2007) Causal Inference in Multisensory Perception. PLoS ONE 2: e943.
24. Dayan P, J Yu A (2003) Uncertainty and learning. IETE Journal of Research 49.2/3: 171–182.
25. Narain D, van Beers RJ, Smeets JBJ, Brenner E (2013) Sensorimotor priors in nonstationary environments. J Neurophysiol. 109: 1259–67. doi: 10.1152/jn.00605.2012.