frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 03 September 2020
doi: 10.3389/frai.2020.00069

A Bayesian Account of Generalist and Specialist Formation Under the Active Inference Framework

Anthony G. Chen[1], *David Benrimoh*[2,3]*, *Thomas Parr*[3] and *Karl J. Friston*[3]

[1] Department of Physiology, McGill University, Montreal, QC, Canada, [2] Department of Psychiatry, McGill University, Montreal, QC, Canada, [3] The Wellcome Centre for Human Neuroimaging, Institute of Neurology, University College London, London, United Kingdom

This paper offers a formal account of policy learning, or habitual behavioral optimization, under the framework of Active Inference. In this setting, habit formation becomes an autodidactic, experience-dependent process, based upon what the agent sees itself doing. We focus on the effect of environmental volatility on habit formation by simulating artificial agents operating in a partially observable Markov decision process. Specifically, we used a "two-step" maze paradigm, in which the agent has to decide whether to go left or right to secure a reward. We observe that in volatile environments with numerous reward locations, the agents learn to adopt a generalist strategy, never forming a strong habitual behavior for any preferred maze direction. Conversely, in conservative or static environments, agents adopt a specialist strategy; forming strong preferences for policies that result in approach to a small number of previously-observed reward locations. The pros and cons of the two strategies are tested and discussed. In general, specialization offers greater benefits, but only when contingencies are conserved over time. We consider the implications of this formal (Active Inference) account of policy learning for understanding the relationship between specialization and habit formation.

Keywords: Bayesian, active inference, generative model, preferences, predictive processing, learning strategies

OPEN ACCESS

Edited by:
Sriraam Natarajan,
The University of Texas at Dallas,
United States

Reviewed by:
Christopher L. Buckley,
University of Sussex, United Kingdom
Fabio Aurelio D'Asaro,
University of Naples Federico II, Italy

****Correspondence:***
David Benrimoh
david.benrimoh@mail.mcgill.ca

Specialty section:
This article was submitted to Machine Learning and Artificial Intelligence,
a section of the journal
Frontiers in Artificial Intelligence

Received: *14 December 2019*
Accepted: *28 July 2020*
Published: *03 September 2020*

Citation:
Chen AG, Benrimoh D, Parr T and Friston KJ (2020) A Bayesian Account of Generalist and Specialist Formation Under the Active Inference Framework. Front. Artif. Intell. 3:69.
doi: 10.3389/frai.2020.00069

INTRODUCTION

Any self-organizing system must adapt to its surroundings if it is to continue existing. On a broad timescale, population characteristics change to better fit the ecological niche, resulting in evolution and speciation (Futuyma and Moreno, 1988). On a shorter timescale, organisms adapt to better exploit their environment through the process of learning. The degree or rate of adaptation is also important. Depending on the environment around the organism, specialization into a specific niche or favoring a more generalist approach can offer distinct advantages and pitfalls (Van Tienderen, 1991). While adopting a single, automatic, behavioral strategy might be optimal for static environments—in which contingencies are conserved—creatures that find themselves in more variable or volatile environments should entertain a broader repertoire of plausible behaviors.

We focus upon adaptation on the shorter timescale in this paper, addressing the issue of behavioral specialization formally within a Markov decision process formulation of Active Inference (Friston et al., 2017). Active inference represents a principled framework in which to describe Bayes optimal behavior. It depends upon the notion that creatures use an internal (generative) model to explain sensory data, and that this model incorporates beliefs about "how

I will behave." Under Active Inference, learning describes the optimization of model parameters—updating one's generative model of the world such that one acts in a more advantageous way in a given environment (Friston et al., 2016). Existing work has focused upon how agents learn the (probabilistic) causal relationships between hidden states of the world that cause sensations which are sampled (Friston et al., 2016, 2017b; Bruineberg et al., 2018; Kaplan and Friston, 2018; Parr and Friston, 2018). In this paper, we extend this formalism to consider learning of policies.

While it is clear that well-functioning agents can update their understanding of the meaning of cues around them—in order to adaptively modulate their behavior—it is also clear that agents can form habitual behaviors. For example, in goal-directed vs. habitual accounts of decision making (Gläscher et al., 2010), agents can either employ an automatic response (e.g., go left because the reward is always on the left) or plan ahead using a model of the world. Habitual responses are less computationally costly than goal-oriented responses; making it desirable to trust habits when they have been historically beneficial (Graybiel, 2008; Keramati et al., 2011). This would explain the effect of practice—as we gain expertise in a given task, the time it takes to complete that task and the subjective experience of planning during the task diminishes, likely because we have learned enough about the structure of the task to discern and learn appropriate habits (Klapp, 1995).

How may our Active Inference agent learn and select habitual behaviors? To answer this question, we introduce a novel feature to the Active Inference framework; namely, the ability to update one's policy space. Technically, a prior probability is specified over a set of plausible policies, each of which represents a sequence of actions through time. Policy learning is the optimization of this probability distribution, and optimizing the structure of this distribution (i.e., "structure learning") through Bayesian model comparison. Habitual behavior may emerge through pruning implausible policies, and reducing the number of behaviors that an agent may engage in. If an agent can account for its behavior without calling on a given policy, it can be pruned, resulting in a reduced policy space, allowing agents to infer which policy it is pursuing more efficiently. Note that in Active Inference, agents have to infer the policy they are pursuing, where this inference is heavily biased by prior beliefs and preferences about the ultimate outcomes. We argue that pruning of redundant behavioral options can account for the phenomenon of specialization (behavior highly adapted to specific environments), and the accompanying loss of flexibility. In addition to introducing Bayesian model reduction for prior beliefs about policies, we consider its biological plausibility, and its relationship with processes that have been associated with structure learning (i.e., the removal of redundant model parameters). Finally, through the use of illustrative simulations, we show how optimizing model structure leads to useful policies, the adaption of an agent to its environment, the effect of the environment on learning and the costs and benefits of specialization. In what follows, we will briefly review the tenets of Active Inference, describe our simulation set up and then review the behavioral phenomenology in light of the questions posed above.

MATERIALS AND METHODS

Active Inference

Under Active Inference, agents act to minimize their variational free energy (Friston, 2012) and select actions that minimizes variational free energy expected following the action. This imperative formalizes the notion that an adaptive agent should act to avoid being in surprising states, should they wish to continue their existence. In this setting, free energy acts as an upper bound on surprise and expected free energy stands in for expected surprise or uncertainty. As an intuitive example, a human sitting comfortably at home should not expect to see an intruder in her kitchen, as this represents a challenge to her continued existence; as such, she will act to ensure that outcomes (i.e., whether or not an intruder is present) match her prior preferences (not being in the presence of an intruder); for example, by locking the door.

More formally, surprise is defined as the negative log probability of observed outcomes under the agent's internal model of the world, where outcomes are generated by hidden states (which the agents have no direct access to, but which cause the outcomes) that depend on the policies which the agent pursues (Parr and Friston, 2017):

$$-\ln P(\tilde{o}) = -\ln\left[\sum_{\tilde{s},\pi} P(\tilde{o}, \tilde{s}, \pi)\right] \quad (1)$$

Here, $\tilde{o} = (o_1, \ldots, o_T)$ and $\tilde{s} = (s_1, \ldots, s_T)$ correspond to outcomes (observations) and states throughout time, respectively, and π represents the policies (sequence of actions through time). Since the summation above is typically intractable, we can instead use free energy as an upper bound on surprise (Friston et al., 2017):

$$F = E_Q[\ln Q(\tilde{s}, \pi) - \ln P(\tilde{o}, \tilde{s}, \pi)] \quad (2)$$

As an agent acts to minimize their free energy, they must also look forward in time and pursue the policy which they expect would best minimize their free energy. The contribution to the expected free energy from a given time, $G(\pi, \tau)$, is the free energy associated with that time, conditioned on the policy, and averaged with respect to a posterior predictive distribution (Friston et al., 2015):

$$G(\pi, \tau) = E_{Q(s_\tau|\pi)P(o_\tau|s_\tau)}\left[\ln P(o_\tau, s_\tau \mid \pi) - \ln Q(s_\tau \mid \pi)\right] \quad (3)$$

We can then sum over all future time-points (i.e., taking the path integral from the current to the final time: $(\pi) = \sum_{t\geq\tau} G(\pi, t)$) to arrive at the total expected free energy expected under each policy.

Partially Observable Markov Decision Process and the Generative Model

A Partially Observable Markov Decision Process (POMDP or MDP for short) is a generative model for modeling discrete hidden states with probabilistic transitions that depend upon a policy. This framework is useful for formalizing planning and decision making problems and has various applications in artificial intelligence and robotics (Kaelbling et al., 1998). An MDP comprises two types of *hidden* variables which the agent must infer: hidden *states* ($\tilde{s}$) and *policies* (π). An MDP agent must then navigate its environment, armed with a generative model that specifies the joint probability distribution of observed outcomes and their hidden causes, and the imperative of minimizing free energy. The states, outcomes and policies are defined more concretely in the following sections.

The MDP implementation consists of the following matrices specifying categorical distributions (Friston et al., 2017b):

$$\begin{aligned}
\mathbf{A_{ij}} &= P\left(o_\tau = i \mid s_\tau = j\right) && \text{state} - \text{outcome mapping}\\
\mathbf{B}\left(\mathbf{u}\right)_{\mathbf{ij}} &= P\left(s_{\tau+1} = i \mid s_\tau = j\,,\right. \\
&\qquad \left. u = \pi\left(\tau\right)\right) && \text{state} - \text{state transition}\\
\mathbf{C}_{\tau,\mathbf{i}} &= P\left(o_\tau = i\right) && \text{outcome preference}\\
\mathbf{D_i} &= P\left(s_1 = i\right) && \text{belief about initial states}\\
\mathbf{E_i} &= P\left(\pi = i \mid E\right) && \text{independent policy prior}
\end{aligned}$$

The generative model (**Figure 1**) assumes that outcomes depend upon states, and that current states depend upon states at the previous timepoint and the action taken (as a result of the policy pursued). Specifically, the state-outcome relationship is captured by an **A** (likelihood) matrix, which maps the conditional probability of any *i-th* outcome given a *j-th* state. A policy, $\pi_i = (u_1, \ldots, u_T)$, is a sequence of actions (u) through time, which the agent can pursue. Generally, an agent is equipped with multiple policies it can pursue. Conceptually, these may be thought of as hypotheses about how to act. As hidden states are inaccessible, the agent must infer its current state from the (inferred) state it was previously in, as well as the policy it is pursuing. State-to-state transitions are described by the **B** (transition) matrix. The **C** matrix encodes prior beliefs about (i.e., a probability distribution over) outcomes, which are synonymous with the agent's preferences. This is because the agent wishes to minimize surprise and therefore will endeavor to attain outcomes that match the distributions in the **C** matrix. The **D** matrix is the prior belief about the agent's initial states (the agent's beliefs about where it starts off). Finally, **E** is a vector of the belief-independent prior over policies (i.e., intrinsic probability of each policy, without considering expected free energy).

A concept that will become important below is *ambiguity*. Assuming an agent is in the *i-th* hidden states, s^i, the probable outcomes are described by a categorical distribution by the *i-th* column of the **A** matrix. We can therefore imagine a scenario where the distribution $P(o_\tau \mid s_\tau = i)$ has *high entropy* (e.g., uniformly distributed), and outcomes are approximately equally likely to be sampled. This is an *ambiguous* outcome. On the other hand, we can have the opposite situation with an *unambiguous* outcome, where the distribution of outcomes given states has *low entropy*. In other words, "if I am in this state, then I will see this and only this." This unambiguous, precise outcome allows the agent to infer the hidden state that they are in.

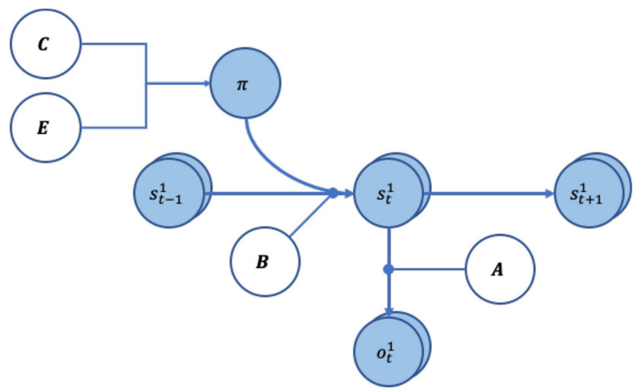


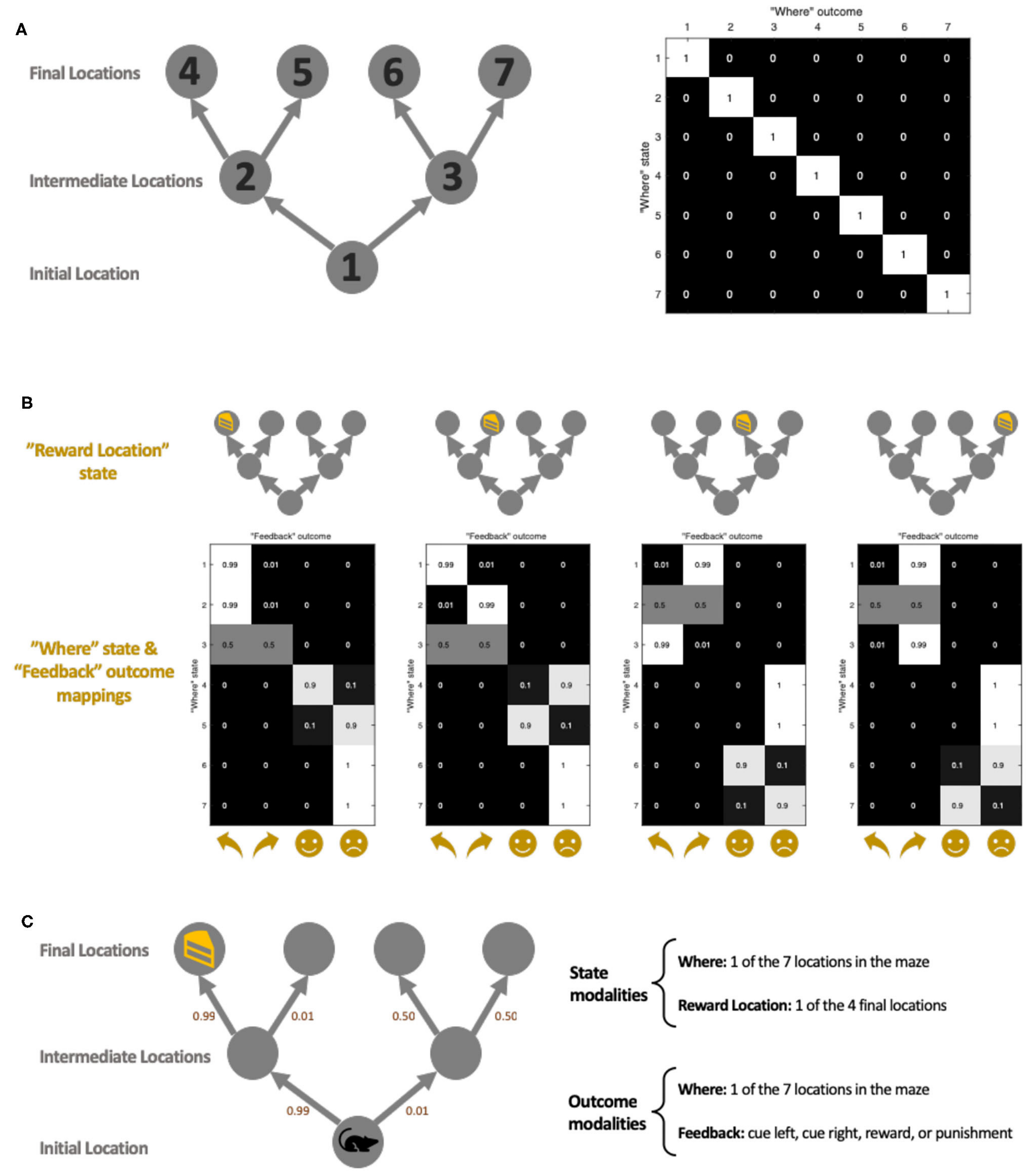
FIGURE 1 | Graphical representation of the generative model. The arrows indicate conditional dependencies, with the endpoint being dependant on where the arrow originated form. The variables in white circles show priors, whereas variables in light blue circles are random variables. The A and B matrices have round arrowhead to show they encode the transition probabilities between the variables.

Crucially, under Active Inference, an agent must also infer which policy it is pursuing at each time step. This is known as planning as inference (Botvinick and Toussaint, 2012). The requisite policy inference takes the form:

$$\pi = \sigma(\hat{\mathbf{E}} - \mathbf{F} - \gamma \cdot \mathbf{G}) \tag{4}$$

Here, π represents a vector of sufficient statistics of the posterior belief about policies: i.e., expectations that each allowable policy is currently in play. ***F*** is the free energy for each policy based on past time points and **G** is the expected free energy for future time points. The free energy scores the evidence that each policy is being pursued, while the expected free energy represents the prior belief that each policy will reduce expected surprise or uncertainty in the future. The expected free energy comprises two parts—*risk* and *ambiguity*. Risk is the difference between predicted and preferred outcomes, while ambiguity ensures that policies are chosen to disclose salient information. These two terms can be rearranged into *epistemic* and *pragmatic* components which, as one might guess, reduce uncertainty about hidden states of the world and maximize the probability of preferred outcomes.

The two quantities required to form posterior beliefs about the best policy (i.e., the free energy and expected free energy of each policy) can be computed using the **A**, **B**, and **C** matrices (Friston et al., 2016; Mirza et al., 2016). The variable γ is an inverse temperature (precision) term capturing confidence in policy selection, and $\hat{\mathbf{E}}$ is the (expected log of the) intrinsic prior probabilities in the absence of any inference (this is covered more in-depth in the "*Policy Learning and Dirichlet Parameters*" section below). The three quantities are passed through a softmax function (which normalizes the exponential of the values to sum to one). The result is the posterior expectation; namely, the most likely policy that the agent believes it is in. This expectation enables the agent to select the action that it thinks is most likely.



"Where" state	1	2	3	4	5	6	7
1	1	0	0	0	0	0	0
2	0	1	0	0	0	0	0
3	0	0	1	0	0	0	0
4	0	0	0	1	0	0	0
5	0	0	0	0	1	0	0
6	0	0	0	0	0	1	0
7	0	0	0	0	0	0	1

"Where" state	↰	↱	☺	☹
1	0.99	0.01	0	0
2	0.99	0.01	0	0
3	0.5	0.5	0	0
4	0	0	0.9	0.1
5	0	0	0.1	0.9
6	0	0	0	1
7	0	0	0	1

"Where" state	↰	↱	☺	☹
1	0.99	0.01	0	0
2	0.01	0.99	0	0
3	0.5	0.5	0	0
4	0	0	0.1	0.9
5	0	0	0.9	0.1
6	0	0	0	1
7	0	0	0	1

"Where" state	↰	↱	☺	☹
1	0.01	0.99	0	0
2	0.5	0.5	0	0
3	0.99	0.01	0	0
4	0	0	0	1
5	0	0	0	1
6	0	0	0.9	0.1
7	0	0	0.1	0.9

"Where" state	↰	↱	☺	☹
1	0.01	0.99	0	0
2	0.5	0.5	0	0
3	0.01	0.99	0	0
4	0	0	0	1
5	0	0	0	1
6	0	0	0.1	0.9
7	0	0	0.9	0.1

State modalities
- **Where:** 1 of the 7 locations in the maze
- **Reward Location:** 1 of the 4 final locations

Outcome modalities
- **Where:** 1 of the 7 locations in the maze
- **Feedback:** cue left, cue right, reward, or punishment

FIGURE 2 | Simulation maze set-up. **(A)** The maze location set-up. There are a total of 7 locations in the maze, each with their corresponding indexes (left diagram). The state-outcome mapping (A matrix) between "Where" (i.e., agent's current location) state and outcome is an identity matrix (right figure), meaning they always

(Continued)

FIGURE 2 | correspond exactly. The maze consists of three stages: initial, intermediate, and final. The state-state transition matrix (B matrix) ensures that an agent can only move forward in the maze, following the direction of the arrow. **(B)** The state-outcome transition probability between the "Where" state and "Feedback" outcome (as encoded by the A matrix). Depending on the location of the reward, the agent receives different feedbacks which include a directional cue (cue left or cue right) in the *initial* and *intermediate locations*, and a reward or punishment at the *final locations*. The index of the y-axis corresponds with the location index in **(A)**. Here we have depicted *unambiguous cues*, where the agent is 99% sure it sees the cue pointed in the correct (i.e., toward the reward location) cue. **(C)** An example maze set-up with a reward at the left-most *final location*. The agent starts in the *initial location*, and the agent's model-based brain contains representations of where it is in the maze, as well as where it thinks the reward is. The agent is able to make geographical observations to see where it is in the maze **(A)**, as well as receive a "feedback" outcome which gives it a cue to go a certain location, or to give it reward/punishment **(B)**. The small numbers beside each arrow illustrate the ambiguity of the cues. As an example, we have illustrated the left-most scenario of **(B)**.

Simulations and Task Set-Up

We return to our question of the effect of the environment on policy learning via setting up a simulated environment in which our synthetic agent (visualized as a mouse) forages (**Figures 2A,C**). Our environment takes the form of a two-step maze inspired by Daw et al. (2011), which is similar to that used in previous work on Active Inference (Friston et al., 2015, 2017). The maze allows for an array of possible policies, and the challenge for our agent is to learn to prioritize these appropriately. The agent has two sets of beliefs about the hidden states of the world: where it is in the maze, and where the reward is. The agent also receives two outcomes modalities: *where* it is in the maze and *feedback* received at each location in the maze (**Figure 2C**, right). The agent always knows exactly where it is in the maze (**Figure 2A**), and receives different "Feedback" outcomes, depending on where it is in the maze and the location of the reward (**Figure 2B**).

The mouse always starts in the same initial location (**Figure 2A**, position 1) and is given no prior information about the location of the reward. This is simulated by setting matrix **D** such that the mouse strongly believes that it is in the "initial location" at $\tau = 1$ but with a uniform distribution over the "reward location." The agent is endowed with a preference for rewarding outcomes and wishes to avoid punishing outcomes (encoded via the **C** matrix). Cues are placed in the initial and intermediate locations (cue left and cue right). While the agent has no preference for the cues *per se*, it can leverage the cue information to make informed decisions about which way to go to receive the reward. In other words, cues offer the opportunity to resolve uncertainty and therefore have salient or epistemic value. **Figure 2C** shows the reward in the left-most final location, accompanied by an *unambiguous cue*—the agent is 99% sure that "cue left" means that the reward is actually on the left. This leads it to the correct reward location. The nature of the maze is such that the agent cannot move backward; i.e., once it reaches the intermediate location it can no longer return to the initial location. Once the agent gets to the final location, it will receive either a reward (if it is at the reward location) or be punished.

To see the effect of training under different environments, we set up two different maze conditions: a *volatile environment*, in which the reward can appear in any one of the 4 final locations with equal frequencies, and a *non-volatile environment*, where the reward only appears on the two left final locations (**Figure 3A**). Crucially, this volatility is between-trial, because these contingencies do not change during the course of a trial. The mouse has no explicit beliefs about changes over multiple trials. Two mice with identical initial parameters are trained in these two distinct environments. With our set-up, each mouse can entertain 7 possible policies (**Figure 3B**). Four of the policies allow the mouse to get to one of the final four locations, whereas three additional policies result in the mouse staying in either the intermediate or initial locations. Finally, both mice are trained for 8 trials per day for 32 days with *unambiguous* cues in the two environments (**Figure 3C**). Bayesian model reduction (further discussed below) is performed in-between training to boost learning. Note that we set-up the training environment with *unambiguous cues* to allow for efficient learning, while the testing environment always has *ambiguous cues*—akin to explicit curriculums of school education vs. the uncertainty of real-life situations.

Policy Learning and Dirichlet Parameters

Whereas inference means optimizing expectations about hidden states given the current model parameters, learning is the optimization of the model parameters themselves (Friston et al., 2016). Within the MDP implementation of Active Inference, the parameters encode sets of categorical distributions that constitute the probabilistic mappings and prior beliefs denoted by **A**, **B**, **C, D,** and **E** above. A Dirichlet prior is placed over these distributions. Since the Dirichlet distribution is the conjugate prior for categorical distributions, we can update our Dirichlet prior with categorical data and arrive at a posterior that is still Dirichlet (FitzGerald et al., 2015).

While all model parameters can be learned (FitzGerald et al., 2015; Friston et al., 2016, 2017b), we focus upon policy learning. The priors are defined as follows:

$$E \sim Dir(e) \tag{5}$$

Here E is the Dirichlet distributed random variable (or parameter) that determines prior beliefs about policies. The variables $e = (e_1, \ldots, e_k)$ are the concentration parameters that parameterize the Dirichlet distribution itself. In the following, k is the number of policies. Policy learning occurs via the accumulation of e concentration parameters—the agent simply counts and aggregates the number of times it performs each policy and this count makes up the e parameters. Concretely, if we define $\pi = (\pi_{\mathbf{1}}, \pi_{\mathbf{2}}, \ldots, \pi_{\mathbf{k}})$ to be the probability the agent observes itself pursuing policies $\pi = 1, \ldots, k$, the posterior distribution over the policy space is:

$$Q(\mathbf{E}) = Dir(\mathbf{e}) = Dir(e + \pi) \tag{6}$$

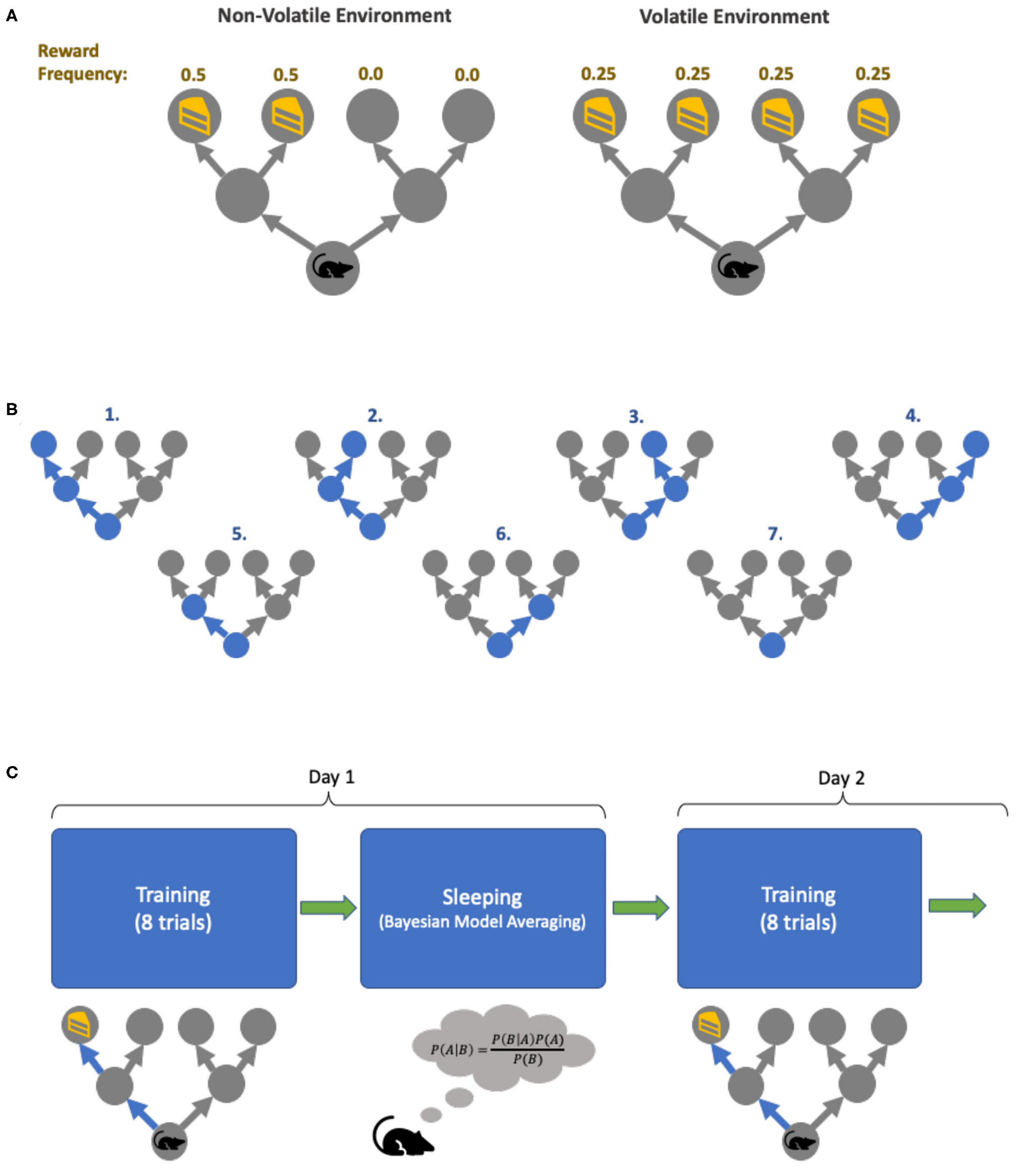


FIGURE 3 | Simulation task set-up. **(A)** The two environments in which the agents are trained. The environment can be *non-volatile* (left), in which the reward always appears on the *left* of the initial location, with equal frequency. The *volatile* environment (right) has reward appearing in all four final locations with equal frequencies. **(B)** The agent's policies. In our simulation, our agents each have 7 policies it can pursue: the first four policies correspond to the agent going to one of the final locations, policies 5–6 has the agent going to one of the intermediate locations and staying there, and policy 7 has the agent not moving from its initial location for the entire duration of a trial. **(C)** The training cycles. Each day, each agent is trained for 8 trials in their respective environment, and in between days the agent goes to "sleep" (and perform Bayesian model averaging to find more optimal policy concentrations). This process is repeated for 32 days.

where $\mathbf{e} = (e_1 + \pi_\mathbf{1}, \ldots, e_k + \pi_\mathbf{k})$ is the posterior concentration parameter. In this way the Dirichlet concentration parameter is often referred to as a "pseudo-count." Intuitively, the higher the e parameter for a given policy, the more likely that policy becomes because more of $Q(\boldsymbol{E})$'s mass becomes concentrated around this policy. Finally, we take the expected logarithm to compute the posterior beliefs about policies in Equation (4):

$$\hat{\mathbf{E}} = \mathbb{E}_{Q(E)} \left[\ln P(\pi | E) \right] \tag{7}$$

The E vector can now be thought of as an empirical prior that accumulates the experience of policies that are carried over from previous trials. In short, it enables the agent to learn about the sorts of things that it does. This experience dependent prior policy enters inference via Equation (4). Before demonstrating this experience dependent learning, we look at another form of learning known variously as Bayesian model selection or structure learning.

Bayesian Model Comparison

In Bayesian model comparison, multiple competing hypotheses (i.e., models or the priors that defines models) are evaluated in relation to existing data and the model evidence for each is compared (Hoeting et al., 1999). Bayesian model averaging (BMA) enables one to use the results of Bayesian model comparison, by taking into account uncertainty about which is the best model. Instead of selecting just the most probable model, BMA allows us to weight models by their relative evidence—to evaluate model parameters that are a weighted average under each model considered. This is especially important in situations where there is no clear winning model (Hoeting et al., 1999).

An organism which harbors alternative models of the world needs to consider its own uncertainty about each model. The most obvious example of this is in the evaluation of different plausible courses of action (policies), each entailing a different sequence of transitions. Such models need to be learnt and optimized (Acuña and Schrater, 2010; FitzGerald et al., 2014) and, rejected, should they fall short. Bayesian model averaging is used implicitly in Active Inference when forming beliefs about hidden states of the world, where each policy is regarded as a model and different posterior beliefs about the trajectory of hidden states under each policy are combined using Bayesian model averaging. However, here, we will be concerned with the Bayesian model averaging over the policies themselves. In other words, the model in this instance becomes the repertoire of policies entertained by an agent.

Returning to our maze task, our artificial agents traverse through the maze each day and aggregate e parameters (Equation 6) to form its daily posterior—that will serve as tomorrow's empirical prior. During Bayesian model reduction, various reduced models are constructed, via strengthening and weakening amalgamations of e parameters. For each configuration of these policy parameters, model evidence is computed, and BMA performed to acquire the optimal posterior, which becomes the prior for the subsequent day. In brief, we evaluated the evidence of models in which each policy's prior concentration parameter was increased by eight, while the remainder were suppressed (by factor of two and four). This creates a model space—over which we can average to obtain the Bayesian model average of concentration parameters in a fast and biologically plausible fashion. Please see Appendix A, section A.1 for a general introduction to Bayesian model reduction and averaging. Appendix A, section A.2 provides an account of the procedures for an example "day." In what follows, we now look at the kinds of behaviors that emerge from day-to-day using this form of autodidactic policy learning—and its augmentation with Bayesian model averaging. We will focus on the behaviors that are elicited in the simulations, while the simulation details are provided in the appropriate figure legends (and open access software—see software note).

RESULTS

Learning

We now turn to our question about the effect of the environment on policy learning. Intuitively, useful policies should acquire a higher e concentration, becoming more likely to be pursued in the future. In simulations, one readily observes that policy learning occurs and is progressive, evident by the increase in e concentration for frequently pursued policies (**Figure 4**), which rapidly reach stable points within 10 days (**Figure 4B**, see **Figure 3C** for the concept of "training days"). Interestingly, the relative policy strengths attain stable points at different levels, depending on the environment in which the agent is trained. In a conservative environment, the two useful policies stabilize at high levels ($e \approx 32$), whereas in a volatile environment, these four useful policies do not reach the same accumulated strengths ($e \approx 25$). Furthermore, the policies that were infrequently used are maintained at lower levels when trained in a non-volatile environment ($e \approx 7$), while they are more likely to be considered for the agent trained in the volatile environment ($e \approx 11$).

We will henceforth refer to the agent trained in the non-volatile environment as the *specialist agent*, and the agent trained in the volatile environment as the *generalist agent.* Anthropomorphically, the specialist agent is, *a priori*, more confident about what to do: since the reward has appeared in the leftward location its entire life, it is confident that it will continue to appear in the left, thus it has predilections for left-going policies (policies 1 and 2 of **Figure 3B**). Conversely, the generalist agent has seen reward appear in multiple locations, thus it experiences a greater level of uncertainty and considers more policies as being useful, even the ones it never uses. We can think of these as being analogous to a general practitioner, who must entertain many possible treatment plans for each patient, compared to a surgeon who is highly skilled at a specific operation.

We can also illustrate the effect of training on the agents' *reward-acquisition rate:* the rate at which the agents successfully arrive at the reward location (**Figure 5**). Here, we tested the agents after each day's training. We see that (**Figure 5B**, left) with just a few days of training, the specialist agent learns the optimal policies and its *reward-acquisition rate* becomes consistently higher than a *naïve agent* with no preference over any of its policies ($e_{naive} = (e_1, \ldots, e_7) = (1, \ldots, 1)$). Conversely,

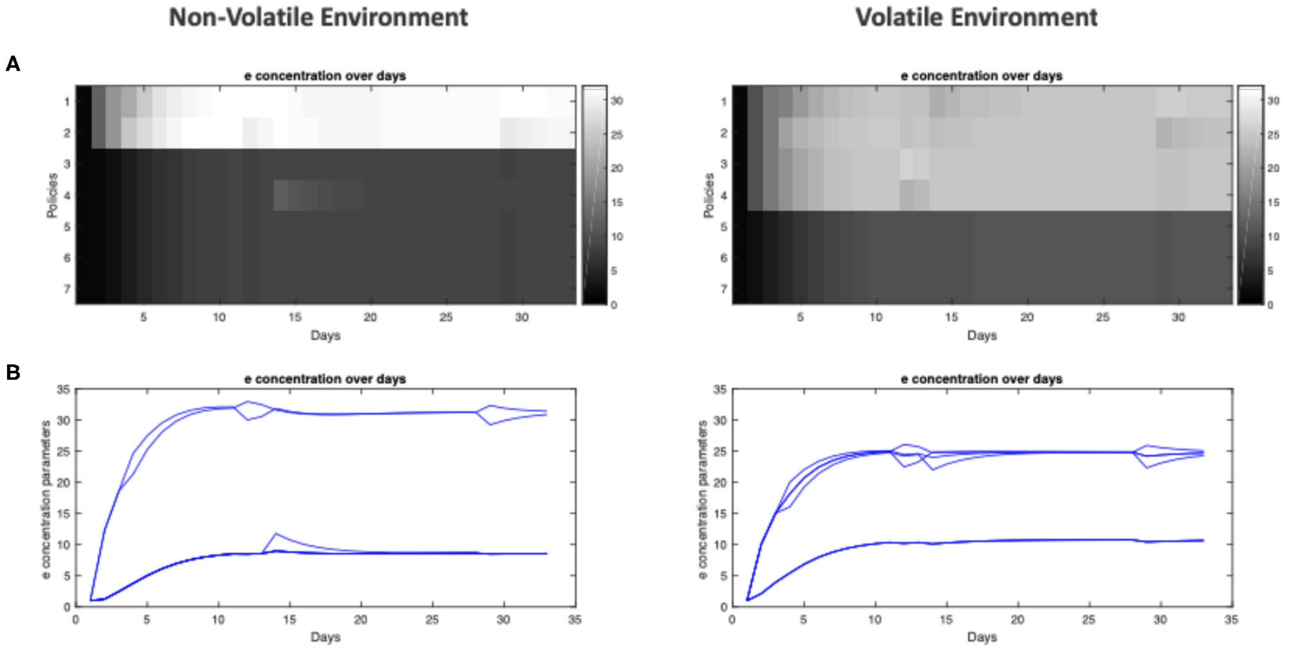


FIGURE 4 | Policy learning over days for agent training in non-volatile and volatile environments. **(A)** Heat-map of *e* concentration parameters for each policy (separated by rows) over all 32 days of training (separated by columns). **(B)** Plot of *e* concentration parameter for policies over 32 days of training.

the generalist agent never becomes an expert in traversing its environment. While it learns to identify the useful policies (**Figure 5A**, right), its performance is never significantly better than the naïve agent (**Figure 5B**, right). We emphasize that the "naive" agent does not simply select policies at random. Rather, it has uninformative policy priors and therefore relies upon its model-based component for policy inference (Equation 4). The similarity in performance between the generalist and "naive" agent is further discussed in the limitations section. Overall, we see that a *non-volatile* environment leads to specialization, whereas a *volatile* environment leads to the agent becoming a generalist.

Testing

We then asked how the specialist and generalist mice perform when transported to different environments. We constructed three testing environments (**Figure 6A**): the *specialized environment,* similar to the environment the specialized agent is trained in; namely, with rewards that only appear on the left side of the starting location (low volatility); the *general environment* containing rewards that may appear in any of the four final locations (high volatility); additionally, the *novel environment* has reward *only* on the right side of the starting location (low volatility).

Each agent was tested for 512 trials in each test environment. Note that the agents do not learn during the testing phase—we simply reset the parameters in our synthetic agents after each testing trial to generate perfect replications of our test settings. We observe that an untrained (naïve) agent has a baseline reward-acquisition rate of ∼60%. On the contrary, the specialist agent excels when the environment is similar to that it trained in, performing at the highest level (89%) out all the agents. In contrast, the specialist agent performs poorly in a general environment (46% reward-acquisition), and fails all but one out of its 512 attempts in a novel environment where it needs to go in the opposite direction to that of its training (**Figures 6B,C**). The generalist agent, being equally trained in all four policies—that take it to one of the end locations—does not suffer from reduced reward-acquisition when exposed to a new environment (the specialized environment or novel environment). However, it does not perform better in a familiar, general environment either. The agent's reward-acquisition remains around 60% across all testing environments, similar to that of a naïve agent (**Figures 6B,C**).

Overall, we find that becoming a specialist vs. a generalist has sensible trade-offs. The benefit of specialization is substantial when operating within the same environment, consistent with data on this topic in a healthcare setting (Harrold et al., 1999; Wu et al., 2001). However, if the underlying environment is different, then performances can decrease to one which is poorer than the performance without specialization.

DISCUSSION

Specialists and Generalists

Our focus in this paper has been on policy optimization, where discrete policies are optimized through learning and Bayesian model reduction. By simulating the development of specialism and generalism, we illustrated the capacity of a generalist to perform in a novel environment, but its failure to reach the level of performance of a specialist in a specific environment. We

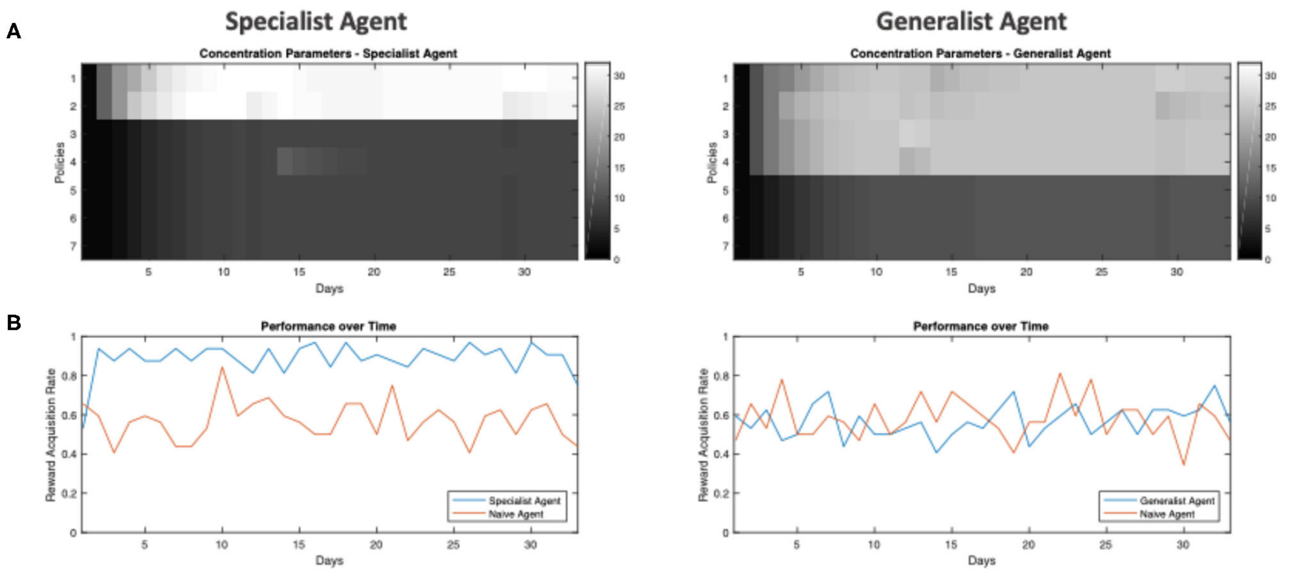


FIGURE 5 | Example performance of in-training agents over days. **(A)** Heat-map of *e* concentration parameters for each policy (separated by rows) over all 32 days of training (separated by columns). **(B)** The frequency at which the agent is able to get to the reward location when tested under ambiguity. This simulated testing is done after each day of training, where each agent is tested under ambiguity (the agent is 65% sure it sees the correct cue) for 32 trials, where the reward location / frequency in the testing environment is identical to the environment in which the agent is trained (i.e., a specialist agent is tested in an environment with low volatility and the reward always being on the left of the initial location). The frequency is computed from how many out of the 32 trials the agent is able to get to the true reward location.

now turn to a discussion of the benefits and costs of expertise. Principally, the drive toward specialization (or expertise) is the result of the organism's imperative to minimize free energy. As free energy is an upper bound on surprise (negative Bayesian model evidence), minimizing free energy maximizes model evidence (Friston et al., 2013). As model evidence takes into account both the accuracy and complexity of an explanation (FitzGerald et al., 2014), it is clear that having a parsimonious model that is well-suited to the environment—a specialist model—will tend to minimize free energy over time, provided the environment does not change.

In a stable (conservative, non-volatile) setting, a complex environment can be distilled down into a simple model without sacrificing accuracy. This results in efficient policy selection and provides a theoretical framework for understanding the formation of expertise. In our simulations, the agent trained in the unchanging environment learns to favor the two policies that go left, as the reward is always on the left of the starting location. It thus becomes more efficient and acts optimally in the face of uncertainty. This is evident by its excellent performance in finding left-situated rewards (**Figure 6**). Indeed, previous theories of expertise differentiate experts from novices in their ability to efficiently generate complex responses to their domain-specific situations (Krampe, 2002; Ericsson, 2008; Furuya and Kinoshita, 2008). For example, in typists, expertise is most well-characterized by the ability to quickly type different letters in succession using different hands (Gentner, 1998; Krampe, 2002). In essence, the expert needs to quickly select from her repertoire of motor policies the most appropriate to type the desired word. This is a non-trivial problem: using just the English alphabet, there are a total of 26^m ways of typing an m-character-long word (e.g., a typist needs to select from $26^6 = 308915776$ policies to type the 6-letter word "EXPERT"). It is no wonder that a beginner typist struggles greatly and needs to forage for information by visually searching the keyboard for the next character after each keystroke. The expert, on the other hand, has an optimized prior over her policy space, and thus is able to efficiently select the correct policies to generate the correct character sequences.

However, specialization does not come without its costs. The price of expertise is reduced flexibility when adapting to new environments, especially when the new settings are contradictory to previous settings (Sternberg and Frensch, 1992; Graybiel, 2008). Theoretically, the expert has a simplified model of their domain, and, throughout their extensive training, has the minimum number of parameters necessary to maintain their model's high accuracy. Consequently, it becomes difficult to fit this model to data in a new, contradictory environment that deviates significantly from the expert's experience. For instance, we observe that people trained in a perceptual learning task perform well in the same task, but perform worse than naïve subjects when the distractor and target set are reversed—and take much longer to re-learn the optimal response than new subjects who were untrained (Shiffrin and Schneider, 1977).

Conversely, a volatile environment precludes specialization. The agent cannot single-mindedly pursue mastery in any particular subset of policies, as doing so would come at the cost of reduced accuracy (and an increase in free energy). The generalist agent therefore never reaches the level of performance that the specialist agent is capable of at its best. Instead, the generalist performs barely above the naïve average reward-acquisition

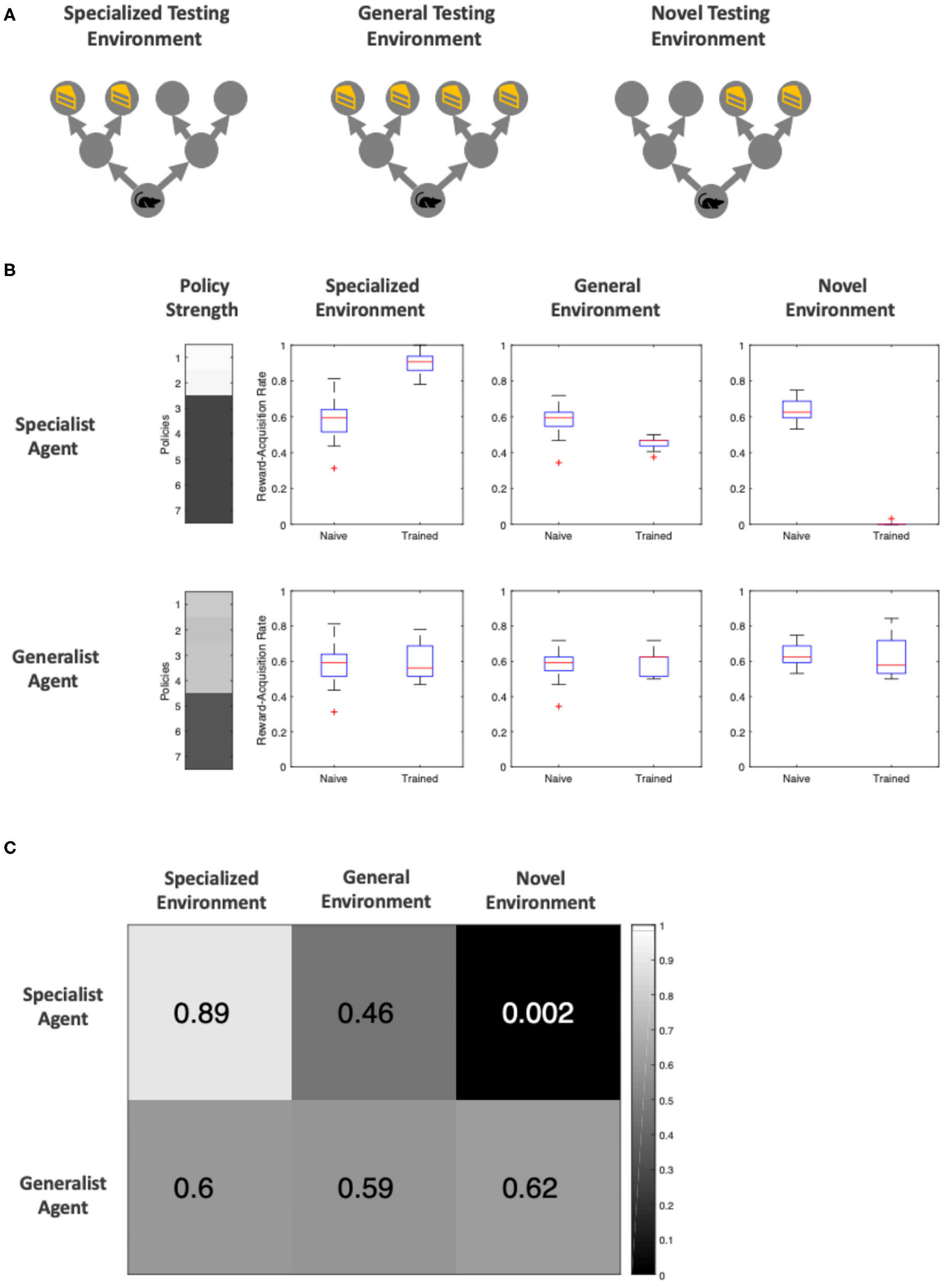


FIGURE 6 | Post-training performance of specialist and generalist agents in ambiguous environments (the agent is 65% sure it sees the cue telling it to go in the correct direction) **(A)** Visualization of the three testing environments. The *specialized* and *general* testing environment have identical reward location and frequencies

(Continued)

FIGURE 6 | top the environments in which the *specialist* and *generalist* agents were trained, respectively. The novel environment is a new, low volatility environment in which the reward only appears to the *right* of the initial location. **(B)** Distribution of reward-acquisition-rate of specialist and generalist agents compared against a naïve agent with no training. The "Policy Strength" column shows how much of each policy the agent has learned, and the three boxes of boxplots show the comparison in performance. The reward-acquisition rate distribution is generated via running each trial 32 times to generate a reward-acquisition rate (proportion of times the agent correctly navigates to the reward location), and repeating this process 16 times to generate a distribution of scores. **(C)** A confusion matrix of mean reward-rate of each agent within each testing environment. Both the heat map and the color over each element represents the reward-acquisition rate.

rate, even when tested under a general environment. However, the generalist is flexible. When placed in novel and changing environments, it performs much better than our specialist agent.

Interestingly, we note that specialist formation requiring a *conservative* training environment adheres to the requirements specified by K. Anders Ericsson in his theory of *deliberate practice*—a framework for any individual to continuously improve until achieving mastery in a particular field (Ericsson et al., 1993, 2009; Ericsson, 2008). Ericsson establishes that deliberate practice requires a well-defined goal with clear feedback (c.f., low volatility learning environment) and ample opportunity for repetition and refinement of one's performance (c.f., training, repetition and, potentially, Bayesian model reduction).

While outside of the current scope, future work could consider even more dynamic (and potentially more realistic) situations where the goal changes intermittently. We tentatively predict if the agent is given time in environments where state-outcome mappings can be inferred easily (unambiguous), it will perform well irrespective of goal location. However, if the environment is always ambiguous, it will be more difficult to learn good habits, and even harder so with an itinerant goal.

Ways of Learning

There are two principal modes of (policy) learning. The first is *learning via reduction,* which entails a naïve agent that starts with an over-complete repertoire of possible policies, who then learns to discard the policies that are not useful. This is how we have tackled policy learning here; specifically, via optimizing a Dirichlet distribution over policies, using Bayesian model reduction. By starting with an abundance of possible policies, we ensure that the best policy is likely to always be present. This also corresponds with the neurobiological findings of childhood peaks in gray matter volume and number of synapses, followed by adolescent decline (Huttenlocher et al., 1982; Huttenlocher and Dabholkar, 1997; Giedd, 2008). In this conceptualization, as children learn they prune away redundant connections, much as our agents triage away redundant policies. Likewise, as the policy spaces are reduced and made more efficient, we also observe a corresponding adolescent decline in brain glucose usage (Chugani et al., 1987). This is consistent with the idea that informational complexity is metabolically more expensive (Landauer, 1961).

The second method of learning is *learning via expansion*. Here, we start with a very simple model and increase its complexity until a more optimal model is reached. Concretely, this problem of increasing a parameter space is one addressed by Bayesian Non-parametric modeling (Ghahramani, 2013), and has been theorized to be utilized biologically for structure learning to infer hidden states and the underlying structures of particular situations (Gershman and Niv, 2010; Collins and Frank, 2013).

Bayesian Model Comparison

In our simulations, we optimized policy strengths through the process of Bayesian model reduction (to evaluate the free energy or model evidence of each reduced model), followed by model averaging—in which we take the weighted average over *all* reduced models. However, BMA is just one way of using model evidences to form a new model. Here, we discuss other approaches to model comparison, their pros and cons, and biological implications. The first is Bayesian model *selection*, in which only the reduced model with the greatest evidence is selected to be the prior for the future, without consideration of competing models. This offers the advantage of reduced computational cost (no need to take the weighted sum during the averaging process) at the cost of a myopic selection—the uncertainty over reduced models is not taken into account.

The second method, which strikes a balance between BMA and Bayesian model selection with respect to the consideration of uncertainty, is BMA with *Occam's Window* (Raftery, 1995). In short, a threshold is established, O_R, and if the log evidence of any reduced model is not within O_R, we simply do not consider that reduced model. Neurobiologically, this would correspond to the effective silencing of a synapse if it falls below a certain strength (Fernando et al., 2012). This way, multiple reduced models and relative uncertainties are still considered, but a great degree of computational cost is saved since less reduced models are considered overall.

We note that in Bayesian model comparison, the repertoire of reduced models to be considered, the width of the Occam's window, as well as the time spent in "wake" (experience-gathering) and "rest" (model comparison and reduction) phases are all hyperparameters. Similar to model parameters, we can expect there to be hyperpriors, which are priors over the hyperparameters. While outside of the scope of the current work, hyperpriors may be optimized via evolutionary processes which also reduce the (path integral of) free energy (Kirchhoff et al., 2018; Linson et al., 2018).

Furthermore, we theorize that there may be a connection between these model optimization processes, and those thought to occur during sleep, in line with previously theorized role of sleep in minimizing model complexity (Hobson and Friston, 2012), and related to the homeostasis hypothesis of sleep (Tononi and Cirelli, 2006). In this theory, a variational free energy minimizing creature tries to optimize a generative model that is both accurate and simple—i.e., that affords the least complicated explanation for the greatest number of observations. Mathematically, this follows from the fact that surprise can be expressed as model evidence—and model evidence is the difference between accuracy and complexity. During wakefulness, an organism constantly receives sensory information, and forms accurate yet potentially complex models to fit these data (neurobiologically, via increases in the number and strength of synaptic connections through associative plasticity). During sleep, which lacks any precise sensory input, creatures can optimize their models *post-hoc* by reducing complexity (Friston and Penny, 2011). This can be achieved by considering reduced (simpler) models and seeing how well they explain the data collected during waking hours (FitzGerald et al., 2014). This is sometimes called Bayesian model reduction (Friston et al., 2018). While we refer to model reduction as "sleep" in this work, we acknowledge that no consensus has been reached on the role of sleep, and the function of sleep as Bayesian model reduction is just one theory.

Computational Psychiatry

Previously, Active Inference has been used as a tool for computational psychiatry, both for phenotyping (Schwartenbeck and Friston, 2016), and as a model of psychiatric symptoms such as illusions (Brown et al., 2013), visual hemineglect (Parr and Friston, 2018), and auditory hallucinations (Benrimoh et al., 2018), to name a few. For instance, low precision assigned to sensory attenuation can result in hallucination (Brown et al., 2013). Uniquely, Active Inference allows for the consideration of both perception and action. Specifically, some recent works have begun to show the potential for disruptions of the policy space to engender symptoms such as visual neglect (Parr and Friston, 2018) and auditory hallucination in schizophrenia (Benrimoh et al., 2018).

While the role of the policy space has been shown to be important, so far, there has been no formal account in Active Inference on how a policy space is learned—in the sense of structure learning—and altered. This is what the current work seeks to provide. Specifically, we formalize the policy to incorporate a policy prior. We then show how this prior is learned, as well as introducing the notion of Bayesian model reduction to change the structure of the policy space. Further, we showcase the interplay between the prior and the free energy in our "two-step" task, where we identified ambiguity—in the state-outcome mapping—as a crucial determinant of when policy priors (i.e., "habits") become important. Depending on the training environment, we demonstrate that different policy priors can underwrite sensible behavior.

Simply put, while we had known that disruption to the policy space plays a role in various psychiatric symptoms, we are now equipped with a formalism to tackle how the policy space can become maladapted to its environment. This can be an experience-dependent process, where rare policies with low priors are never considered. This may also be a result of model-comparison, where the models compared may not have full support over the policy space, or the model averaging process may not consider the full set of possible policies (e.g., due to computational constraints). These are tentative hypotheses, which future work can explore in greater depth.

Moreover, we have focused on ease-of-interpretability in this work and hope this paper can also act as a foundational "tutorial" for future work in Active Inference that seeks to investigate the interaction between the policy space and behavior. We have therefore refrained from making claims about specific brain areas. One can note that policies are usually associated with the striatum (Parr and Friston, 2018), while observation space is modality dependent, per the functional anatomy of primary and secondary sensory cortex (for instance, the state-outcome mapping in auditory tasks can be tentatively theorized to map to the Wernicke's—prefrontal connection). For more precise process theories on how the Active Inference machinery maps onto brain areas, we invite the readers to look at the discussion sections of Benrimoh et al. (2018) and Parr and Friston (2018).

Limitations

One limitation of our simulations was that our agents did not learn about cues at the same time they were learning about policies; in fact, the agents were constructed with priors on which actions were likely to lead to rewards, given specific cues (that is, a correctly perceived cue-left was believed by the agents to—and actually did—always lead to a reward on the left). As such, we did not model the learning of cue-outcome associations and how these may interact with habit formation. We argue this is a reasonable approximation to real behavior; where an animal or human first learns how cues are related to outcomes, and, once they have correctly derived a model of environmental contingencies, can then proceed to optimizing policy selection.

Additionally, while we were able to see a significant performance difference between specialist and generalist agents, there was little distinction between the performance of generalist and naïve agents. This likely resulted from the "two-step" maze being a relatively simple task. As agents are incentivized to go to the very end of the maze to receive a reward, the naïve agents and the generalist agents (as a result of the volatile training environment) have isomorphic prior beliefs about the final reward locations, and thus perform similarly. In this sense, becoming a generalist is the process of resisting specialization, and the preservation of naivety.

To address the above limitations, future work could involve more complex tasks to more clearly differentiate between specialist, generalist, and naïve agents. Additional types of learning should also be included, such as the learning of state-outcome mappings [optimizing the model parameters of the likelihood (**A**) matrix, as described in Friston et al. (2016, 2017b)], to understand how learning of different contingencies influence one another. In addition, more complex tasks may afford the opportunity to examine the generalization of specialist knowledge to new domains (Barnett and Ceci, 2002). This topic has recently attracted a great deal of attention from the artificial intelligence community (Pan and Yang, 2010; Hassabis et al., 2017).

Furthermore, it would be interesting to look at policy learning using a hierarchical generative model, as considered for deep temporal models (Friston et al., 2017a). This likely leads to a more accurate account of expertise-formation, as familiarity with a domain-specific task should occur at multiple-levels of the neural-computation hierarchy (e.g., from lower level "muscle memory" to higher level planning). Likewise, more unique cases of learning can also be explored, such as the ability and flexibility to re-learn different tasks after specializing, and different ways of conducting model comparison (as discussed above).

CONCLUSION

In conclusion, we have presented a computational model under the theoretical framework of Active Inference that equips an agent with the machinery to learn habitual policies via a prior probability distribution over its policy space. In our simulations, we found that agents who specialize—employing a restricted set of policies because these were adaptive in their training environment—can perform well under ambiguity but only if the environment is similar to its training experiences. On the contrary, a generalist agent can more easily adapt to changing, ambiguous environments, but is never as successful as a specialist agent in a conservative environment. These findings cohere with the previous literature on expertise formation—as well as with common human experience. Finally, these findings may be important in understanding aberrant inference and learning in neuropsychiatric diseases.

DATA AVAILABILITY STATEMENT

All simulation scripts used for this article can be found on GitHub (https://github.com/im-ant/ActiveInference_PolicyLearning). Simulation is constructed using the MATLAB package SPM12 (https://www.fil.ion.ucl.ac.uk/spm/). Specifically, the DEM toolbox in SPM12 is used to run the Active Inference simulations.

AUTHOR CONTRIBUTIONS

DB, TP, and KF conceptualized the project and helped supervise it. AC was involved in the investigation along with DB. AC did the data curation, formal analysis, visualization, and writing of the first draft, while receiving methodology help from TP and KF. AC, DB, TP, and KF took part in the review and edits of the subsequent drafts. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by Rosetrees Trust (Award Number 173346) to TP. KF was a Wellcome Principal Research Fellow (Ref: 088130/Z/09/Z).

ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at Chen et al. (2019).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2020.00069/full#supplementary-material

REFERENCES

Acuña, D. E., and Schrater, P. (2010). Structure learning in human sequential decision-making. *PLoS Comput. Biol.* 6:e1001003. doi: 10.1371/journal.pcbi.1001003

Barnett, S. M., and Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychol. Bull.* 128, 612–637. doi: 10.1037/0033-2909.128.4.612

Benrimoh, D., Parr, T., Vincent, P., Adams, R. A., and Friston, K. (2018). Active inference and auditory hallucinations. *Comput. Psychiatry*. 2, 183–204. doi: 10.1162/cpsy_a_00022

Botvinick, M., and Toussaint, M. (2012). Planning as inference. *Trends Cogn. Sci.* 16, 485–488. doi: 10.1016/j.tics.2012.08.006

Brown, H., Adams, R. A, and Parees, I., Edwards, M., and Friston, K. (2013). Active inference, sensory attenuation and illusions. *Cognit. Process.* 14, 411–427.

Bruineberg, J., Rietveld, E., Parr, T., van Maanen, L., and Friston, K. J. (2018). Free-energy minimization in joint agent-environment systems: a niche construction perspective. *J. Theor. Biol.* 455, 161–178. doi: 10.1016/j.jtbi.2018.07.002

Chen, A. G., Benrimoh, D., Parr, T., and Friston, K. J. (2019). A Bayesian account of generalist and specialist formation under the active inference framework. *bioRxiv [Preprint]*. doi: 10.1101/644807

Chugani, H. T., Phelps, M. E., and Mazziotta, J. C. (1987). Positron emission tomography study of human brain functional development. *Ann. Neurol.* 22, 487–497. doi: 10.1002/ana.410220408

Collins, A. G. E., and Frank, M. J. (2013). Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychol. Rev.* 120, 190–229. doi: 10.1037/a0030852

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

Ericsson, K. A. (2008). Deliberate practice and acquisition of expert performance: a general overview. *Acad. Emerg. Med.* 15, 988–994. doi: 10.1111/j.1553-2712.2008.00227.x

Ericsson, K. A., Krampe, R. T., and Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychol. Rev.* 100, 363–406. doi: 10.1037/0033-295X.100.3.363

Ericsson, K. A., Nandagopal, K., and Roring, R. W. (2009). Toward a science of exceptional achievement: attaining superior performance through deliberate practice. *Ann. N. Y. Acad. Sci.* 1172, 199–217. doi: 10.1196/annals.1393.001

Fernando, C., Szathmáry, E., and Husbands, P. (2012). Selectionist and evolutionary approaches to brain function: a critical appraisal. *Front. Comput. Neurosci.* 6:24. doi: 10.3389/fncom.2012.00024

FitzGerald, T. H. B., Dolan, R. J., and Friston, K. (2015). Dopamine, reward learning, and active inference. *Front. Comput. Neurosci.* 9:136. doi: 10.3389/fncom.2015.00136

FitzGerald, T. H. B., Dolan, R. J., and Friston, K. J. (2014). Model averaging, optimal inference, and habit formation. *Front. Hum. Neurosci.* 8, 1–11. doi: 10.3389/fnhum.2014.00457

Friston, K. (2012). A free energy principle for biological systems. *Entropy* 14, 2100–2121. doi: 10.3390/e14112100

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O. Doherty, J., and Pezzulo, G. (2016). Active inference and learning. *Neurosci. Biobehav. Rev.* 68, 862–879. doi: 10.1016/j.neubiorev.2016.06.022

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., and Pezzulo, G. (2017). Active inference: a process theory. *Neural Comput*. 29, 1–49. doi: 10.1162/NECO_a_00912

Friston, K., Parr, T., and Zeidman, P. (2018). Bayesian model reduction. *arXiv [Preprint]* arXiv:1805.07092.

Friston, K., and Penny, W. (2011). Post hoc Bayesian model selection. *Neuroimage* 56, 2089–2099. doi: 10.1016/j.neuroimage.2011.03.062

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., and Pezzulo, G. (2015). Active inference and epistemic value. *Cogn. Neurosci.* 6, 187–224. doi: 10.1080/17588928.2015.1020053

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., and Dolan, R. J. (2013). The anatomy of choice: active inference and agency. *Front. Hum. Neurosci*. 7, 1–18. doi: 10.3389/fnhum.2013.00598

Friston, K. J., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., and Ondobaka, S. (2017b). Active inference, curiosity and insight. *Neural Comput*. 29, 2633–2683. doi: 10.1162/neco_a_00999

Friston, K. J., Rosch, R., Parr, T., Price, C., and Bowman, H. (2017a). Deep temporal models and active inference. *Neurosci. Biobehav. Rev*. 77, 388–402. doi: 10.1016/j.neubiorev.2017.04.009

Furuya, S., and Kinoshita, H. (2008). Expertise-dependent modulation of muscular and non-muscular torques in multi-joint arm movements during piano keystroke. *Neuroscience* 156, 390–402. doi: 10.1016/j.neuroscience.2008.07.028

Futuyma, D. J., and Moreno, G. (1988). The evolution of ecological specialization. *Annu. Rev. Ecol. Syst*. 19, 207–233. doi: 10.1146/annurev.es.19.110188.001231

Gentner, D. R. (1998). "Chapter 1: Expertise in typewriting," in *The Nature of Expertise*, eds M. T. H. Chi, R. Glaser, and M. J. Farr (Taylor & Francis Group), 1–21. doi: 10.4324/9781315799681

Gershman, S. J., and Niv, Y. (2010). Learning latent structure: Carving nature at its joints. *Curr. Opin. Neurobiol*. 20, 251–256. doi: 10.1016/j.conb.2010.02.008

Ghahramani, Z. (2013). Bayesian non-parametrics and the probabilistic approach to modelling. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci*. 371:20110553. doi: 10.1098/rsta.2011.0553

Giedd, J. N. (2008). The teen brain: insights from neuroimaging. *J. Adolesc. Health* 42, 335–343. doi: 10.1016/j.jadohealth.2008.01.007

Gläscher, J., Daw, N., Dayan, P., and O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* 66, 585–595. doi: 10.1016/j.neuron.2010.04.016

Graybiel, A. M. (2008). Habits, rituals, and the evaluative brain. *Annu. Rev. Neurosci.* doi: 10.1146/annurev.neuro.29.051605.112851

Harrold, L. R., Field, T. S., and Gurwitz, J. H. (1999). Knowledge, patterns of care, and outcomes of care for generalists and specialists. *J. Gen. Intern. Med*. 14, 499–511. doi: 10.1046/j.1525-1497.1999.08168.x

Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron* 95, 245–258. doi: 10.1016/j.neuron.2017.06.011

Hobson, J. A., and Friston, K. J. (2012). Waking and dreaming consciousness: neurobiological and functional considerations. *Prog. Neurobiol*. 98, 82–98. doi: 10.1016/j.pneurobio.2012.05.003

Hoeting, J., Madigan, D., Raftery, A., and Volunsky, C. (1999). Bayesian model averaging: a tutorial. *Stat. Sci.* 14, 382–401.

Huttenlocher, P. R., and Dabholkar, A. S. (1997). Regional differences in synaptogenesis in human cerebral cortex. *J. Comp. Neurol*. 387, 167–178. doi: 10.1002/SICI1096-986119971020387:2<167::AID-CNE1>3.0.CO;2-Z

Huttenlocher, P. R., de Courten, C., Garey, L. J., and Van der Loos, H. (1982). Synaptogenesis in human visual cortex–evidence for synapse elimination during normal development. *Neurosci. Lett*. 33, 247–252. doi: 10.1016/0304-3940(82)90379-2

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artif. Intell*. 101, 99–134. doi: 10.1016/S0004-3702(98)00023-X

Kaplan, R., and Friston, K. J. (2018). Planning and navigation as active inference. *Biol. Cybern*. 112, 323–343. doi: 10.1007/s00422-018-0753-2

Keramati, M., Dezfouli, A., and Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Comput. Biol*. 7:e1002055. doi: 10.1371/journal.pcbi.1002055

Kirchhoff, M., Parr, T., Palacios, E., Friston, K., and Kiverstein, J. (2018). The markov blankets of life: autonomy, active inference and the free energy principle. *J. R. Soc. Interface* 15:20170792. doi: 10.1098/rsif.2017.0792

Klapp, S. T. (1995). Motor response programming during simple choice reaction time: the role of practice. *J. Exp. Psychol. Hum. Percept. Perform*. 21, 1015–1027. doi: 10.1037/0096-1523.21.5.1015

Krampe, R. T. (2002). Aging, expertise and fine motor movement. *Neurosci. Biobehav. Rev*. 26, 769–776. doi: 10.1016/S0149-7634(02)00064-7

Landauer, R. (1961). Irreversibility and heat generation in the computing process. *IBM J. Res. Dev*. 5, 183–191. doi: 10.1147/rd.53.0183

Linson, A., Clark, A., Ramamoorthy, S., and Friston, K. (2018). The active inference approach to ecological perception: general information dynamics for natural and artificial embodied cognition. *Front. Robot A. I.* 5:21. doi: 10.3389/frobt.2018.00021

Mirza, M. B., Adams, R. A., Mathys, C. D., and Friston, K. J. (2016). Scene construction, visual foraging, and active inference. *Front. Comput. Neurosci*. 10:56. doi: 10.3389/fncom.2016.00056

Pan, S. J., and Yang, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng*. 22, 1345–1359. doi: 10.1109/TKDE.2009.191

Parr, T., and Friston, K. J. (2017). Working memory, attention, and salience in active inference. *Sci. Rep*. 7, 1–21. doi: 10.1038/s41598-017-15249-0

Parr, T., and Friston, K. J. (2018). The computational anatomy of visual neglect. *Cereb. Cortex*. 28, 777–790. doi: 10.1093/cercor/bhx316

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociol. Methodol.* 25, 111–163.

Schwartenbeck, P., and Friston, K. (2016). Computational phenotyping in psychiatry: a worked example. *eNeuro* 3, 1–18. doi: 10.1523/ENEURO.0049-16.2016

Shiffrin, R. M., and Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychol. Rev*. 84, 127–190. doi: 10.1037/0033-295X.84.2.127

Sternberg, R. J., and Frensch, P. A. (1992). "On being an expert: a cost-benefit analysis," in *The Psychology of Expertise* (New York, NY: Springer New York), 191–203.

Tononi, G., and Cirelli, C. (2006). Sleep function and synaptic homeostasis. *Sleep Med. Rev*. 10, 49–62. doi: 10.1016/j.smrv.2005.05.002

Van Tienderen, P. H. (1991). Evolution of generalists and specialists in spatially heterogeneous environments. *Evolution* 45, 1317–1331. doi: 10.1111/j.1558-5646.1991.tb02638.x

Wu, A. W., Young, Y., Skinner, E. A., Diette, G. B., Huber, M., Peres, A., et al. (2001). Quality of care and outcomes of adults with asthma treated by specialists and generalists in managed care. *Arch. Intern. Med*. 161, 2554–2560. doi: 10.1001/archinte.161.21.2554

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.